
Analytical and computational methods towards a metabolic model of ageing in *Caenorhabditis elegans*

A thesis submitted for the degree of Doctor of Philosophy

submitted by
Florian Geier

Section of Computational and Systemic Medicine
Department of Surgery & Cancer
Faculty of Medicine
Imperial College London

London, 2013

Supervisors: Dr. Jake G. Bundy, Dr. Elizabeth E. Want, Dr. Armand M. Leroi
Examiners: Prof. Howard R. Morris, Prof. Royston Goodacre
Day of submission: September 19, 2013
Word count: 67,963 (excluding references and appendix)

Abstract

Human life expectancy is increasing globally. This has major socioeconomic implications, but also raises scientific questions about the biological bases of ageing and longevity. Research on appropriate model organisms, such as the nematode worm *Caenorhabditis elegans*, is a key component of answering these questions. Ageing is a complex phenomenon, with both environmental and genetic influences. Metabolomics, the analysis of all small molecules within a biological system, offers the ability to integrate these complex factors to help understand the role of metabolism in ageing.

This thesis addresses the current lack of methods for *C. elegans* metabolite analysis, with a particular focus on combining analytical and computational approaches.

As a first essential step, *C. elegans* metabolite extraction protocols for NMR, GC-MS and LC-MS based analysis were optimized. Several methods to improve the coverage, automatic annotation and data analysis steps of NMR and GC-MS are proposed.

Next, stable isotope labelling was explored as a tool for *C. elegans* metabolomics. An automated stable isotope based workflow was developed, which identifies all biological, non-redundant features within a LC-MS acquisition and annotates them with molecular compositions. This demonstrated that the vast majority (> 99.5%) of detected features inside LC-MS metabolomics experiments are not of biological origin or redundant.

This stable isotope workflow was then used to compare the metabolism of 24 different *C. elegans* mutant strains from different pathways (e.g. insulin signalling, TOR pathway, neuronal signalling), with differing levels of lifespan extension compared to wild-type worms. The biologically relevant features (metabolites) were detected and annotated, and compared across the mutants. Some metabolites were correlated with longevity across the mutant set, in particular, glycerophospholipids. This led to the formulation of a hypothesis, that lifespan extension in *C. elegans* requires increased activity of common downstream longevity effector mechanisms (autophagy, and mitochondrial biogenesis), that also involve subcellular compartmentation and hence membrane formation. This results in the alterations in lipid metabolism detected here.

Copyright

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution-Non Commercial-No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or distribution, researchers must make clear to others the license terms of this work.

Statement of originality

Hereby I certify, that this thesis and all the research it describes are the product of my own work, unless specifically stated. All people who contributed to this thesis in any way, will be acknowledged on the next page. As much work of this thesis was based on interdisciplinary collaborations, those collaborations, including a description of my involvement can be found on page 9. However, contributing collaborators will also be acknowledged in the according experimental sections of this thesis.

Acknowledgements

First, I would like to thank my supervisors. Dr Jake Bundy, Professor Armand Leroi and Dr Elizabeth Want. Thank you for this incredible opportunity and giving me all the freedom to work on my own ideas, but providing me with every possible help, whenever I got stuck. Jake, thank you for all the patience with the very details of this work and Armand, thanks for always taking me back to the big picture. I don't only want to thank you for the academic support, but also some great times and discussions outside the college. Thank you.

I very much would like to acknowledge my sponsor, the AXA Research Fund, who actually did so much more than just enable this PhD financially. Dear AXA team and Godefroy Beauvallet in particular, thank you for the opportunities, to make my research accessible to the public. I very much enjoyed the annual workshops in Paris and the great scientific dialogue. Many thanks to the AXA insurance group and Henri de Castries, for a prime example of corporate philanthropy, by establishing (and actually continuing) the AXA Research Fund.

Also, I want to thank all the societies and sponsors for my scholarships and travel awards I received throughout my academic career.

I also want thank Prof. Jeremy Nicholson and Prof. Elaine Holmes for providing this department's excellent, but very accessible analytical infrastructure.

Often unacknowledged, but I really want to thank Katia, Mat, Charlie², Karen, Wendy, (Nahid!), Julia, and all the others who actually keep the department running on a daily basis, so that we scientists can focus on our research. The same - on a college scale - applies to Ian, "Dave" and the whole facilities, security, gym, VWR store and cafeteria team (and all of Imperial's "behind the scenes" teams). Thank you for your incredible job!

I also want to thank all the academics and members of our department. Particularly Dr Timothy Ebbels and Dr Kirill Veselkov for advice on statistics and Dr Richard Barton for his all-round scientific expertise. I thank Dr Olaf Beckonert for his selfless help with any general lab or NMR related issue! Dr Pete Simpson on his advice with ct-HSQC pulse sequences. I also want to mention, that Prof. Nigel Gooderham has been very supportive and always good to talk to since my first day at Imperial. Dr Jake Pearce for making sure the IT worked and helping me to do an emergency hard drive swap. Also, thank you Sarah, for being very patient whilst getting me started with the worms.

A big thank you to Dr Manuel Liebeke and Dr Volker Behrends who weren't only helpful in solving many scientific and day-to-day issues, but also good friends!

I really want to thank all my collaborators for helping me, exchanging your knowledge and actually a lot of fun! Thanks to Dr Xavier Casadevall I Solvas (Xevi) for all the amazing work, patience and hours of cleanroom fiddling to make our microfluidic "worm sorter" work. Thanks to Dr Sukana Zeitoun for her help with the Metallothionein KO experiments. Thanks to Dr Sarah Fearn for helping me with the ToF-SIMS acquisitions. I very much want to thank Yevgeniy Izraelit for the great time we had during his visit and all the projects we kicked off together. I also want to thank Yevgeniy's PI, Prof. Frank Schroeder, for his all his input into our projects, interesting scientific debates and the amazing opportunity to visit Cornell University for a few days and present my work there.

I specifically want to thank my industrial collaborators from Waters Corp. in Manchester: Dr Steve Pringle and Dr Martin Palmer, who helped me to acquire the data for my stable isotope annotation algorithm. This was such a great experience and opportunity! Also many thanks to John Shockcor for great (scientific) debates, particularly at ASMS 2011.

Now to the people whose company kept me from boarding the next plane and aborting this PhD. My friends.

Two key people who I cannot thank enough are Paul Benton and Matthew Lewis. Both, in their role of collaborators on several great projects, sources of great knowledge, but mostly as very very good friends. Thank you guys! - I couldn't have done this without you.

Kathy Lewis - thank you. For all the fun times as well as tougher ones and most of all for the patience with some occasional drama. I am very glad you and Mat came to London.

Abdullah, thank you for "Hotel Alkandari" - Such a lovely place, plenty of room, anytime of the year. I also thank you for teaching the German the correct use of "however" and "respectively". And so much more..

Gabriel, best (master) student I ever had. I am very proud. Thanks you not only for all the hard work, but also being a very good friend and the helpful input into my work.

I would like to write a paragraph about everyone, but unfortunately there is a word limit (which by feeling I have already exceeded). Therefore thanks to the rest of the crew - for great discussions on science, unrelated issues and all the fun times: Alex, Anas, Anna, Cancy, Claire, Elisa, Esti, Evie, Lea, Maria, Nicole, Panos, Perrine, Renaud and Steve.

Also thanks to my MRes friends Eugene, Gurpreet, Jen, Jess, Sarah and Tom. It was good to find such great friends straight from my first day at Imperial.

I was lucky enough to also have had many friends outside the department. For example from my times at the board of GermanSoc: Marie, Rolf, Lorenz and Natalie. Thanks for all the help and fun.

Also, I want to thank some very close friends, I have been lucky to have since my second masters project. Christina, Tina and Dylan.

Daniel and Julian - thanks for keeping me reminded that there is a life after the thesis.

Also, I want to thank all my bachelor, master and project students I had - not only for the great work they delivered, but also for the (continuing) friendship. I hope the occasional free meals and drinks made up for making you work very hard. Thanks Gabriel (again!), Angevil, Sabrina, Guo-Heng, Hugo, Amy, Alex and Simon.

I must also thank my friends, moral supporters and occasional welcomed distractors back in Munich, Germany - particular during writeup in Munich. Carl, Diana, Elisabeth, Giulia, Julius, Kathy, Teresa, Saline and Korbi & "the boys". I particularly want to thank Claire Andonian and her lovely family, especially for two amazing weeks in New York.

Also, I really want to acknowledge my former employer, Prof. Neuburger and the whole Institute of Actuary Mathematics - where I learnt many of the IT and programming skills without which this project could have never succeeded.

Many thanks to my great undergraduate professors, particularly professor Bracher, Friess, Vollmar and Winter - through who I managed to get into Imperial in the first place.

To my "UK family". Dear uncle Michael, dear Alice thank you for welcoming me to the UK and being like a second home and a link to English culture. The same applies to Sandra, Tom, Tim, Kurt and Natalie. Thank you for all the priceless experiences.

Now to the most important people. My family. Thanks to my brother, Benedict, for all the moral support, help with biology, welcomed distraction and a great month during his visit at Imperial. Thanks to my parents, Margit and Wolfgang and my grandparents, particularly my granddad - my secret hero. Without your continuous support, motivation and financial assistance this work would have never been completed. Neither any of my other academic degrees. Thank you.

Many thanks to all of you! In case someone, who actually reads this thesis and feels unacknowledged, despite having had a significant contribution to this work: I shall buy you a drink.

Collaborative projects

	Project		
1	The microfluidic worm sorter		
2	Metabolomics of oxidative stress responses <i>C. elegans</i>		
3	ToF-SIMS imaging of <i>C. elegans</i>		
4	Automatic NMR HSQC metabolite annotation		
5	LC-NMR ion pair separation of <i>C. elegans</i> extracts		
6	Profiling of long and short lived mitochondria KO mutants		
7	Data reduction and annotation using stable isotopes		
8	LC-MS analysis of long-lived <i>C. elegans</i> mutants		
	Collaborator	Collaborating Lab	University / Company
1	Xavier Casadevall I Solvas	Andrew deMello	Imperial College London
2	Sukaina Zeitoun-Ghandour	Stephen Stuerzenbaum	King's College London
3	Sarah Fearn	David McPhail	Imperial College London
4	Yevgeniy Izraelit	Frank Schroeder	Cornell University
5	Yevgeniy Izraelit	Frank Schroeder	Cornell University
6	Marija Sumakovic	Aleksandra Trifunovic	University of Cologne
7	Steve Pringle, Martin Palmer	-	Waters Corp.
8	Matthew Lewis	-	Imperial College London
	Publication or thesis chapter	My task	
1	[84]	Wormculture, experiments (partially)	
2	[660]	Sample preparation, NMR, data analysis	
3	[185]	Experiment planning, sample preparation, analysis	
4	III.2	Sample generation, data analysis	
5	III.3	LC-NMR, data analysis	
6	III.4	Sample prep., GC-MS, data analysis	
7	IV.	Everything, but wet-lab LC-MS	
8	V.	Everything, but wet-lab LC-MS	

Metabolomics *versus* Metabonomics

This work is mainly based on global metabolic profiling, also referred to as "metabonomics" [444] or "metabolomics" [455, 171]. However, "In practice, the terms are often used interchangeably, and the analytical and modelling procedures are the same".

Furthermore, "The distinction is mainly philosophical, rather than technical." [443]

In order to be consistent, across this thesis, the term metabolomics will be used.

Contents

I.. Introduction	24
I.1. Ageing and longevity	24
I.2. <i>C. elegans</i> as model organism	26
I.3. Systems biology and metabolomics	29
I.4. Metabolomics of ageing and <i>C. celegans</i>	32
I.5. Aims and structure of this thesis	33
II.. Evaluation of <i>C. elegans</i> metabolite extraction strategies for untargeted cross-platform profiling	36
II.1. Introduction	37
II.2. Design of Experiment	38
II.3. Materials and Methods	40
II.3.1. Sampling and extraction	40
II.3.2. Analytics	42
II.3.2.1. NMR	42
II.3.2.2. GC-MS	43
II.3.2.3. LC-MS	43
II.4. Results	45
II.4.1. Contribution of bacterial biomass to total signal	45
II.4.2. Assigned metabolites	45
II.4.3. Distribution of LC-MS features	48
II.4.4. Multivariate clustering and similarity	51
II.4.5. Reproducibility	51
II.4.6. Yield	55
II.5. Discussion	58
II.5.1. Necessity of a washing step during sampling	58
II.5.2. What is the "best" method	58
II.5.3. Observed metabolites - In comparison to previously published studies	59
II.5.4. What is limiting the coverage and assignment?	60
III.Methods to increase the metabolite coverage and annotation on NMR and GC-MS based profiling platforms	62
III.1.Introduction	63
III.2.HSQC sensitivity improvement by universal ¹³ C labelling	64
III.2.1. Background and introduction	64

III.2.2. Materials and Methods	66
III.2.2.1. Sample generation	66
III.2.2.2. Acquisition and processing	66
III.2.2.3. Spectral deconvolution and automatic database matching	67
III.2.3. Results and Discussion	69
III.2.3.1. ^{13}C enrichment can achieve a drastic increase in sensitivity	69
III.2.3.2. The constant time version outperforms the regular HSQC pulse sequence for ^{13}C enriched sample types.	69
III.2.3.3. ^{13}C enriched ct-HSQC spectra are ideal for automatic compound annotation	73
III.3. Online and offline LC-NMR separation of <i>C. elegans</i> extracts	79
III.3.1. Background and introduction	79
III.3.2. Materials and Methods	81
III.3.2.1. Sample generation	81
III.3.2.2. Instrumentation and acquisition	81
III.3.3. Results and Discussion	84
III.3.3.1. The ion pair reagent needs to be deuterated	84
III.3.3.2. Online LC-NMR sensitivity is too low for the <i>C. elegans</i> extract	86
III.3.3.3. Offline ion pair chromatography compliments reversed phase	88
III.4. Workflow for GC-MS TIC alignment for easy visual interpretation and application to longevity mutants	91
III.4.1. Background and introduction	91
III.4.2. Materials and Methods	93
III.4.2.1. Sample generation, extraction and acquisition	93
III.4.2.2. Data processing	93
III.4.2.2.1. Retention time alignment (for TIC overlay)	94
III.4.2.2.2. Standard GC-MS processing workflow	95
III.4.2.2.3. Processing of differential features, not found by the standard workflow	95
III.4.3. Results and Discussion	97
III.4.3.1. Workflow for the GC-MS retention time alignment for TIC overlay	97
III.4.3.2. TIC alignment generates interesting hypotheses for the <i>C. elegans</i> mitochondrial mutant dataset	99
III.5. Conclusion	103
IV.A tool for data reduction and annotation of LC-MS features	104
IV.1. Introduction	105
IV.1.1. Motivation and Background	105
IV.1.1.1. Data reduction in untargeted LC-MS profiling	105
IV.1.1.2. Determination of a feature's molecular formula	106
IV.1.2. This experiment outline	108

IV.2. Methods	110
IV.2.1. Sample generation and metabolite extraction	110
IV.2.1.1. Preparation of a modular base minimal medium	110
IV.2.1.2. Determining the sulphur requirement of <i>E. Coli</i> on the modified minimal medium	111
IV.2.1.3. Generation of stable isotope labelled <i>E. coli</i> samples . . .	111
IV.2.1.4. Generation of stable isotope labelled <i>C. elegans</i> samples .	112
IV.2.1.5. Metabolite extraction	113
IV.2.2. LC-MS data acquisition	113
IV.2.3. LC-MS data processing	115
IV.2.4. Programming and simulations	115
IV.3. Results & Discussion	117
IV.3.1. <i>In silico</i> considerations for enhanced molecular formula calculations by isotope label analysis	118
IV.3.1.1. Knowing the carbon, nitrogen and sulphur atom count al- lows unambiguous annotation at ToF mass accuracy . . .	118
IV.3.1.2. Impact of molecular weight for unambiguous molecular for- mula calculation (considering different atom-count constraints)	123
IV.3.2. A protocol for stable isotope labeling of carbon, nitrogen and sulphur in <i>C. elegans</i>	126
IV.3.3. Development of a software algorithm for automated data reduction and molecular formula generation by stable isotopes	128
IV.3.4. Optimisation and testing of the data reduction and formula genera- tion algorithm	136
IV.3.4.1. Testing and optimising the data reduction algorithm . . .	136
IV.3.4.2. Testing the annotation by the formula generator	142
IV.4. Conclusion	151
V.. A metabolic model of longevity in <i>C. elegans</i>	152
V.1. Introduction	153
V.1.1. Motivation	153
V.1.2. Selection of long-lived mutant strains	154
V.1.3. Longevity promoting genes in <i>C. elegans</i>	156
V.1.3.1. Insulin like signalling (ILS)	156
V.1.3.2. Protein Turnover, autophagy and dietary restriction medi- ated by target of rapamycin (TOR)	158
V.1.3.3. Sirtuins (SIR)	161
V.1.3.4. c-Jun N-terminal kinases (JNK)	162
V.1.3.5. Hypoxia-induced factor (HIF-1)	162
V.1.3.6. Mitochondria	163
V.1.3.7. Super oxide dismutase (SOD) and reactive oxygen species (ROS)	164
V.1.3.8. Fertility and germline cell signalling (GCS)	165

V.1.3.9. Sensory perception and neuronal control	166
V.1.4. Outline of experiments	168
V.2. Methods	170
V.2.1. Sample generation and metabolite extraction	170
V.2.1.1. Generation of longevity mutant screen samples	170
V.2.1.2. Generation of stable isotope annotation samples	171
V.2.1.3. Metabolite extraction	171
V.2.2. LC-MS acquisition	172
V.2.3. Processing and Analysis	173
V.2.3.1. Processing of the isotope annotation samples	173
V.2.3.2. Database matching	173
V.2.3.3. Processing of the mutant screen samples	174
V.2.3.4. Data Analysis	177
V.3. Results	180
V.3.1. Data reduction and annotation of <i>C. elegans</i> features	180
V.3.2. Database matching of <i>C. elegans</i> adduct/formula pairs	184
V.3.3. Processing and matching of the mutant dataset	186
V.3.3.1. Optimising the peak detection parameters	186
V.3.3.2. Alignment of the m/z domain	187
V.3.3.3. Dataset matching and retention time alignment	187
V.3.3.4. Preparing the data matrix for analysis	191
V.3.4. Data analysis - the role of metabolism in <i>C. elegans</i> longevity	193
V.3.4.1. A single, all strain model, is not the best data analysis strategy	193
V.3.4.2. Most mutant's metabolic phenotype differs from the control	195
V.3.4.3. The multivariate link between life expectancy and metabolism	197
V.3.4.4. Inter-strain comparison shows a distinct, systematic metabo- lite signature	200
V.3.4.5. The inter-strain metabolite signature is correlated to life expectancy	203
V.3.4.6. Particular phospholipid classes correlate with life expectancy	205
V.4. Discussion	210
V.4.1. A fair compromise between targeted and untargeted	210
V.4.2. An increased phospholipid pool links <i>C. elegans</i> longevity pathways	211
V.4.3. The next steps	214
V.5. Conclusion	217
VI. General discussion, future technologies and perspectives	218
VI.1. Summary of the key results	218
VI.2. Two key problems and two promising key technologies	221
VI.2.1. The problem of measuring ageing: "The worm sorter"	221
VI.2.2. Biology is spatial: TOF-SIMS	224

VI.3. Current issues	227
VI.3.1. Metabolomics	227
VI.3.2. (<i>C. elegans</i>) ageing research	228
VI.4. Conclusion	229

Appendices 287

Theory and methods 288

A1. The basic constituents of matter	288
A2. Analytical chemistry	289
A2.1. Separation science	290
A2.1.1. Basic principles of chromatography	290
A2.1.2. Liquid chromatography	291
A2.1.2.1. Instrumentation	291
A2.1.2.2. Common modes of liquid chromatography	292
A2.1.3. Gas chromatography	294
A2.1.3.1. Instrumentation	294
A2.1.3.2. Derivatisation	295
A2.1.3.3. Retention time indices and retention time locking	296
A2.2. Analyte detection	297
A2.2.1. Nuclear magnetic resonance (NMR) spectroscopy	297
A2.2.1.1. The quantum physics behind NMR	297
A2.2.1.2. The vector model of NMR spectroscopy	299
A2.2.1.3. The chemical shift	301
A2.2.1.4. Spin-Spin coupling	302
A2.2.1.5. NMR for quantification	305
A2.2.1.6. NMR instrumentation	305
A2.2.1.7. NMR sample set acquisitions	308
A2.2.1.8. Two dimensional NMR spectroscopy	310
A2.2.2. Mass spectrometry	311
A2.2.2.1. Types of ionisation sources	312
A2.2.2.2. Types of mass analysers	313
A2.2.2.3. The Detector	320
A2.2.2.4. Fragmentation and hybrid instruments: The QQQ and Q-ToF	320
A2.3. Hyphenated approaches	322
A2.3.1. LC-NMR	322
A2.3.2. Liquid chromatography coupled to mass spectrometry (LC-MS)	324
A2.3.2.1. Electrospray ionisation	324
A2.3.2.2. LC-MS sample set acquisitions	327
A2.3.2.3. Different types of chromatograms: TIC, BIP, EIC	328

A2.3.3.	Gas chromatography coupled to mass spectrometry (GC-MS)	328
A2.3.3.1.	Electron impact ionisation (EI)	330
A2.3.3.2.	GC-MS sample set acquisitions	333
A2.4.	How to choose an analytical platform	334
A3.	Informatics	337
A3.1.	Data processing	337
A3.1.1.	Processing of NMR spectra	338
A3.1.1.1.	Window functions, phase and baseline correction and shift referencing	338
A3.1.1.2.	Alignment versus fitting or binning	340
A3.1.1.3.	Processing of 2D NMR spectra	341
A3.1.2.	Processing of LC-NMR spectra	341
A3.1.3.	Processing of LC-MS data	341
A3.1.3.1.	The LC-MS deconvolution and alignment work- flow: XCMS	342
A3.1.4.	Processing of GC-MS data	346
A3.1.4.1.	Fragmentation patterns and database matching	346
A3.1.4.2.	GC-MS processing workflow	349
A3.2.	Data analysis and statistics	350
A3.2.1.	Univariate statistics	350
A3.2.1.1.	Distributions and descriptive statistics	350
A3.2.1.2.	Statistical inference	353
A3.2.2.	Data preprocessing: Normalisation, centering and scaling	357
A3.2.3.	Multivariate statistics and pattern recognition	360
A3.2.3.1.	Principal component analysis (PCA)	361
A3.2.3.2.	Partial least squares regression (PLS)	365
A3.2.3.3.	Partial least squares - discriminant analysis (PLS- DA)	366
Protocols		367
B1.	Wormculture	367
B1.1.	Growth medium	367
B1.2.	Age synchronisation	368
B2.	Default sampling and extraction protocol	368
B3.	NMR spectroscopy	369
B3.1.	Preparation	369
B3.2.	Acquisition	369
B3.3.	Processing and analysis	370
B4.	Gas Chromatography - Mass Spectrometry	370
B4.1.	Preparation	370
B4.2.	Acquisition	370
B4.3.	Processing	371
B4.4.	Analysis	371

B5. Liquid Chromatography - Mass Spectrometry	371
Supplementary tables and figures	372
C1. Supplementary tables and figures for chapter II	372
C2. Supplementary tables and figures for chapter III	374
C3. Supplementary tables and figures for chapter IV	383
C4. Supplementary tables and figures for chapter V	386

List of Abbreviations

2AA 2-Aminoadipate

2KG 2-Ketoglutarate

BIP Base ion peak chromatogram

CGC *C. elegans* genetics center (Madison, Wisconsin, USA)

ct – HSQC constant time Heteronuclear Single Quantum Coherence

CV Coefficient of variation

DAF Dauer affection factor

diAcPC Diacylglycerophosphocholine

EIC Extracted ion chromatogram

FA Fatty acid

GC Gas chromatography

HIF Hypoxia inducible factor

HMDB Human Metabolome Database

HPLC High performance liquid chromatography

HSQC Heteronuclear Single Quantum Coherence

ILS Insulin-like signalling

LC Liquid chromatography

monoAcPC Monoacylglycerophosphocholine

MS Mass spectrometry

NMR Nuclear magnetic resonance

PC Phosphatidylcholine

<i>PCA</i>	Principal component analysis
Q^2	Cross validated coefficient of determination
R^2	Coefficient of determination
<i>SD</i>	Standard deviation
<i>SN</i>	Signal to noise (ratio)
<i>TIC</i>	Total ion chromatogram
<i>TOF</i>	Time of flight
<i>TOR</i>	Target of rapamycin
<i>UPLC</i>	Ultra performance liquid chromatography

List of Figures

I.1.	Changing demography	24
I.2.	<i>C. elegans</i>	26
I.3.	<i>C. elegans</i> lifecycle	28
I.4.	System biology's OMICS	30
I.5.	Workflow of a typical LC-MS experiment	31
I.6.	Overview of thesis structure	34
II.1.	Design of experiment	38
II.2.	Bacterial "contamination" of <i>C. elegans</i> pellet	46
II.3.	Distribution of LC-MS metabolite features	49
II.4.	LC-MS BIP chromatograms of M80 and MCW extraction	50
II.5.	Principal component analysis of extracted metabolite features	52
II.6.	Precision of extracted metabolite features	54
II.7.	Yield of extracted metabolite features	57
III.1.	Natural abundance vs ¹³ C enriched (ct) - HSQC	70
III.2.	HSQC vs ct- HSQC detailed	71
III.3.	Effect of the constant time HSQC sequence on C-C couplings	72
III.4.	Peaks detected in the ¹² C and ¹³ C (constant time) HSQC spectra	74
III.5.	Monte Carlo simulation to find the ideal matching error window	76
III.6.	HSQC cross peaks matched to HMDB metabolites	77
III.7.	Spectral overlap between <i>C. elegans</i> extract and ion pair reagent	85
III.8.	Online LC-NMR experiments	87
III.9.	Fractions from offline LC-NMR experiments	89
III.10.	Outline of the proposed (TIC) GC-MS alignment workflow	97
III.11.	TIC before and after retention time alignment	98
IV.1.	Outline of the chapter	108
IV.2.	Instrument accuracy requirements impacted by various atom count constraints I	120
IV.3.	Instrument accuracy requirements impacted by various atom count constraints II	121
IV.4.	Required accuracy to determine molecular composition versus its mass	124
IV.5.	Minimum sulphur requirement for <i>E. coli</i> strain NCM3722	127
IV.6.	Isotopic abundance pattern of (in)completely labelled lipid	128
IV.7.	Workflow of data reduction and annotation algorithm	130

IV.8.	Check tables for isotope peak pairs	133
IV.9.	Multiply charged and isotope pair mismatch	134
IV.10.	Optimising the retention time and m/z window during matching	137
IV.11.	Ideal rtE and ppmE settings for the four datasets	139
IV.12.	Different rtE and ppmE settings for the four datasets	140
IV.13.	Adduct distributions in the test data sets	143
IV.14.	Molecular formulae of HILIC ESI+	145
IV.15.	Molecular formulae of reversed phase ESI+	146
IV.16.	Molecular formulae of HILIC ESI-	147
IV.17.	Molecular formulae of reversed phase ESI-	148
V.1.	Pathways of lifespan regulation in <i>C. elegans</i>	157
V.2.	Design of experiment	169
V.3.	<i>C. elegans</i> BIP with detected metabolite candidates highlighted	181
V.4.	Adducts detected in the <i>C. elegans</i> extract	182
V.5.	Molecular formulae of reversed phase ESI+	183
V.6.	Database matches for generated formula/adduct pairs	185
V.7.	Determining the matching window for the target peaks within the longevity mutant sample set I	188
V.8.	Retention time alignment based on grouped, target peaks	190
V.9.	Determining the matching window for the target peaks within the longevity mutant sample set II	191
V.10.	Intensity distribution before and after intensity normalisation	192
V.11.	PCA and PLS-DA models including all samples	194
V.12.	PLS regression models of life expectancy	198
V.13.	Predicted versus actual lifespan extension	199
V.14.	Clustered heat maps of (significant) fold changes	202
V.15.	Univariate intensity - lifespan correlation for each metabolite	204
V.16.	Intensity fold change and life expectancy correlation	205
V.17.	Different lipid classes determine life expectancy	207
V.18.	Chain length, degree of saturation and correlation with life expectancy of mono- and diacylglycerophosphocholines	208
V.19.	Proposed model of longevity through increased compartmentalisation	214
VI.1.	The "Worm sorter" may solve the "offspring problem"	223
VI.2.	ToF-SIMS imaging mass spectrometry	226
A2.1.	Precision and accuracy	289
A2.2.	Liquid chromatography (LC) system	292
A2.3.	Gas Chromatograph	295
A2.4.	Spin physics of NMR spectroscopy	298
A2.5.	Vector model of NMR spectroscopy	300
A2.6.	¹ H NMR chemical shift range of different functional groups	301

A2.7.	NMR spectrum of ethanol and the corresponding spin populations	303
A2.8.	Chemical shift and coupling in the NMR vector model	304
A2.9.	The 180° refocusing pulse	306
A2.10.	Assembly of a NMR magnet and probe	307
A2.11.	Standard procedure of a NMR sample acquisition	308
A2.12.	Mass spectrum of ethanol	311
A2.13.	Basic assembly of a mass spectrometer	312
A2.14.	Quadrupole mass analyser	316
A2.15.	Time-of-flight (ToF) mass analyser	318
A2.16.	Q-ToF (quadrupole time-of-flight) mass analyser	322
A2.17.	On-flow LC-NMR system	323
A2.18.	Electrospray ionisation (ESI) mechanism	325
A2.19.	Different types of chromatograms (TIC,BIP,EIC)	329
A2.20.	Mass spectrum of peak apex from EIC in figure A2.19c	330
A2.21.	Electron impact ionisation (EI) source	331
A2.22.	GC-EI-MS fragmentation	332
A2.23.	Standard procedure of a GC-MS sample acquisition	333
A2.24.	Choosing an appropriate analytical platform	335
A3.1.	NMR spectra processing workflow	339
A3.2.	Wavelet peak detection in xcms	343
A3.3.	Retention time alignment in xcms	345
A3.4.	GC-MS spectrum	347
A3.5.	The normal distribution	351
A3.6.	Probability and its distribution	354
A3.7.	Effect of mean centering and scaling on data	359
A3.8.	Principal component analysis (PCA)	363
A3.9.	PCA scores and loadings plot interpretation	364
C2.1.	Aligned spectra of the mitochondrial mutant dataset	382
C3.1.	Accuracy requirement for <i>ab initio</i> formula calculation, impacted by various atom-count constraints	385
C4.1.	LOESS function to correct post-acquisition calibrate m/z domain	386
C4.2.	ROC curves for optimising xcms peak detection parameters	387
C4.3.	Hierarchical clustered heat map of all samples	388
C4.4.	Glycerophosphocholines versus glycerophosphoethanolamines	389

List of Tables

II.1. Putative compounds detected via GC-MS and NMR	47
III.1. Reversed phase gradient	82
III.2. Ion pairing gradient	82
III.3. <i>C. elegans</i> strains used to test the GC-MS TIC alignment	92
IV.1. Trace element solution (1000x) for modified Gutnick medium	110
IV.2. Base medium	111
IV.3. Stable isotope label combinations	112
IV.4. Gradient of reversed phase acquisition	114
IV.5. Gradient of HILIC acquisition	114
IV.6. Dataset for algorithm development and validation	129
IV.7. Exact mass and difference (considering mass defect) of used isotopes	131
IV.8. Reduction algorithms's performance benchmark	141
IV.9. Number of elemental compositions calculated for the test datasets	142
V.1. Longevity Mutant Strains	155
V.2. LC-MS RP gradient for final <i>C. elegans</i> metabolite annotation and longevity mutant screen	172
V.3. Database matches for generated formula/adduct pairs	186
V.4. PCA and PLS-DA models of mutant versus control	196
V.5. Correlating the model quality and lifespan extension	197
A2.1. Basic properties of different mass analysers. [619, 130] Exact values depend on the instrumental setup and vary between instrument manufacturers. . .	314
A3.1. Typical number of data points and file size of different analytical platforms	337
A3.2. Contingency table of statistical hypothesis testing outcomes	356
A3.3. Data preprocessing methods	360
C1.1. Compounds detected via GC-MS and NMR	373
C2.1. HMDB metabolites matched with ¹² C and ¹³ C HSQC cross peaks	374
C3.1. Minimum and maximum atom-count to m/z ratios	383
C3.2. Minimum and maximum elemental ratios	383
C3.3. Adducts used during molecular formula calculation	384

I.. Introduction

I.1. Ageing and longevity

Even though life expectancy has steadily risen over the last thousands of years, the rapid increase during the last century was nothing like before (Figure I.1).

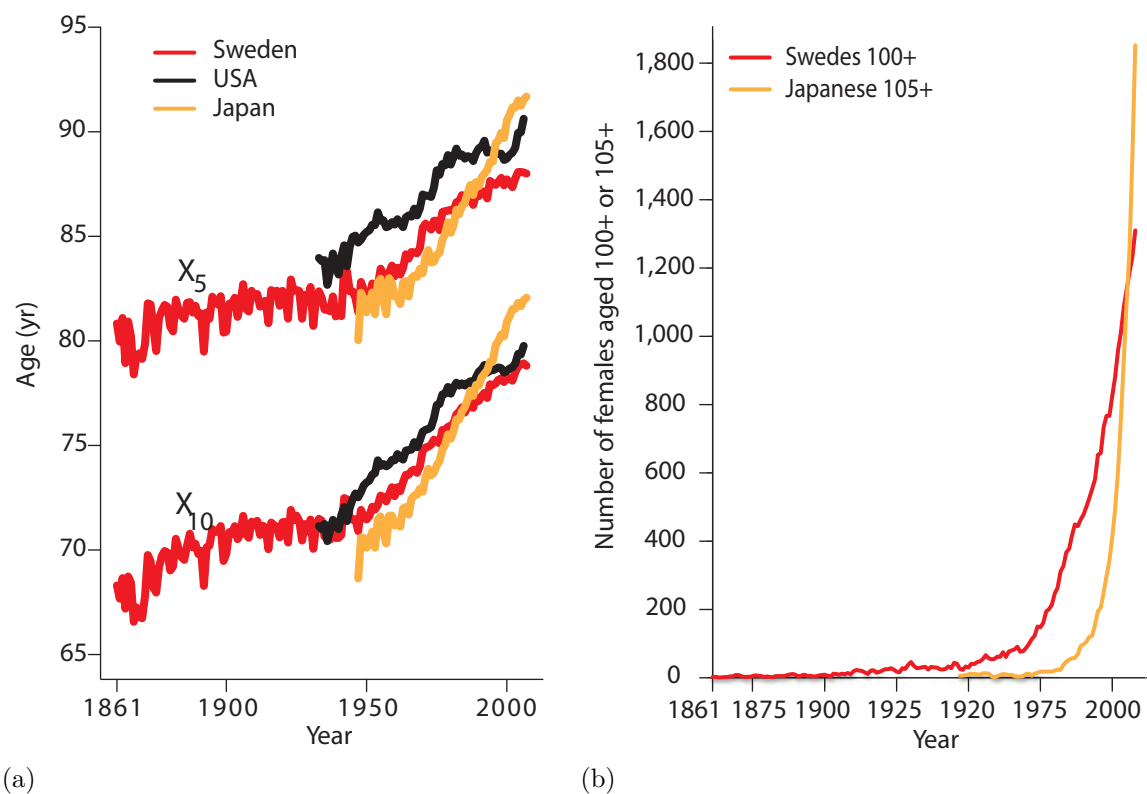


Figure I.1.: **Changing demography:**

Life expectancy displayed a steep increase over the past century. X_5 and X_{10} in figure (a) indicate the ages at which 5 and 10 years life expectancy remain, respectively, and their increase over time. Figure (b) shows the drastic increase in centenarians - people over 100 years. (The figures were adapted from [594])

Thirty years of additional life expectancy or more were gained in several countries throughout the last century [97]. Currently life expectancy increases with a pace of 2.5 years per decade, or 6 hours a day in wealthy countries [593]. If this trend continues, then most children born in these countries since 2000 will turn 100 [97]. These demographic changes will have a far reaching economic impact, not only for federal social security system, but also insurance companies. Inevitably, rising expenses for pensions and healthcare will need to be faced, being exacerbated by a world-wide decrease of birth rate in western countries [219]. Interestingly, the constant offset between 5 and 10 year life expectancy in figure I.1a suggests that senescence is postponed, rather than decelerated [594].

The causes of senescence, *i.e.* biological ageing, are still unknown, but many theories have been proposed. In general one has to differentiate between mechanistic and evolutionary explanations [325, 263]. Important mechanistic theories include the somatic mutation, telomere loss, mitohermesis, free radical [230] and protein waste accumulation, theories. Established evolutionary explanations include the mutation accumulation [407] and the antagonistic pleiotropy theory [627]. However, a recent point of view is to acknowledge that none of the theories can be seen as sufficient [326]. They need to be interlinked and complement each other, which leads to the definition of network theories of ageing [327].

Genetics has a limited influence on an individual's life-span. In their natural habitats, most animals will not reach their maximum possible age. This is due to challenging environmental conditions, such as availability of nutrition, succumbing to predators, a harsh climate, or other reasons. In a laboratory setting, however, under controlled and optimal environmental conditions, maximum life-span is mainly limited by genetic predisposition [287, 124]. The very same trend has been observed for humans. Demographic studies in twin cohorts showed, that only 25 % of the variance of life expectancy can be explained by genetic disposition [401, 240]. This value though, may slightly increase at higher age [595]. For the oldest of a population, however, genetics seems to be the primary driver of life-expectancy [475]. This explains the prior mentioned, drastic increase in population life-span: Throughout the last century, we have tamed our environmental conditions by medical progress, understanding of hygiene, balanced and secure food supply and implementing social security systems. Thus, more people survive to an age where genetic disposition governs the remaining lifespan. This shows that an understanding of the genetics behind ageing becomes increasingly important. However, the fact that overall lifespan still

increases with such a pace, also demonstrates that many life-expectancy influencing environmental factors still remain to be elucidated or to be improved.

Apart from demographic studies, such as above, humans themselves are rarely subject to ageing research, in order to understand those factors and general limits of longevity. Apart from ethical issues, this is because of their long absolute life expectancy and the impact of non-controllable, confounding factors, such as nutrition, medication and other lifestyle choices. For those reasons, model systems, in controlled environments, are used [309]. They range from simple systems such as bacteria [659], fungi (yeast) [290], cultured cells, over multicellular invertebrates [307, 308, 465] like fruit flies [208], nematode worms [183], up to fish [191] and even mammals such as mice [573] or dogs [611].

The model system used in this thesis is the nematode worm *Caenorhabditis elegans*.

I.2. *C. elegans* as model organism

Caenorhabditis elegans, shown in figure I.2 is a soil inhabiting nematode of approximately 1 mm length, feeding on bacteria. Being a nematode, it is a multicellular animal, containing around one thousand cells. In fact, the hermaphrodite has exactly 959 and the male 1031 cells.



Figure I.2.: **Adult *C. elegans*:**
3.5 days after hatching (at 20 °C), a hermaphrodite *C. elegans* reaches adulthood and measures 1.1 mm in length and 30 μm in radius.

It was first established as a model organism by Sydney Brenner in 1974 [60]. It was soon adopted by many research labs around the globe, [149], making it one of the best understood organisms. It was the first multicellular animal to have its 97 million base pair genome completely sequenced [79]. Over 35 % of its genes have homologues in humans, with whom it shares many conserved pathways [99, 522]. Thus, *C. elegans* research may have direct implications for human biology.

C. elegans is also convenient to handle. It can be grown on agar plates, feeding on an *E. coli* lawn, with one plate supporting up to ten thousand individuals. It can be frozen, stored, and thawed when needed. As it is transparent, its cellular development can be observed by light microscopy. In fact, each of its cell's fates can be traced and has been mapped in detail, making it an important model for developmental biology [150]. As hermaphrodites can reproduce asexually, their offspring are genetic clones. This and the ease of silencing specific genes, using RNA interference (RNAi) [223] through feeding of siRNA expressing bacteria [294], also made it an indispensable system for genetics. A full genome RNAi library is available [293] and thousands of genetically modified mutant strains can readily be ordered from the CGC (Caenorhabditis Genetics Center) [239]. The "worm community" also provides a plethora of other resources, such as the WormBook (wormbook.org), a peer reviewed collection of *C. elegans* biology, methods and protocols and Wormbase (wormbase.org), which is a large bioinformatics repository [654].

One reason which made *C. elegans* an indispensable model for development and ageing is its short, but precisely timed lifecycle [78], which is summarised in figure I.3. Depending on the temperature, which may be varied between 15°C and 25°C, a worm's lifecycle from hatching to egg laying takes between 3 and 5 days. As they have an exoskeleton (cuticle), composed of cross-linked collagen, they undergo molting. Four phases of linear growth, with subsequent moulting [331], define *C. elegans*'s larval stages, from L1 to L4. After the final molt, the young adult stage, spermatogenesis is complete and oogenesis begins. Once completed, they are considered adults, which starts at the onset of egg laying. A self-fertilising hermaphrodite can lay up to 300 eggs, but if fertilised by a male, this may increase to well over 1000 eggs [244]. Even though fertility ceases after a few days, the (wild type) worm may live on up to two more weeks.

As seen in figure I.3, a worm may also take a different lifecycle. The morphologically

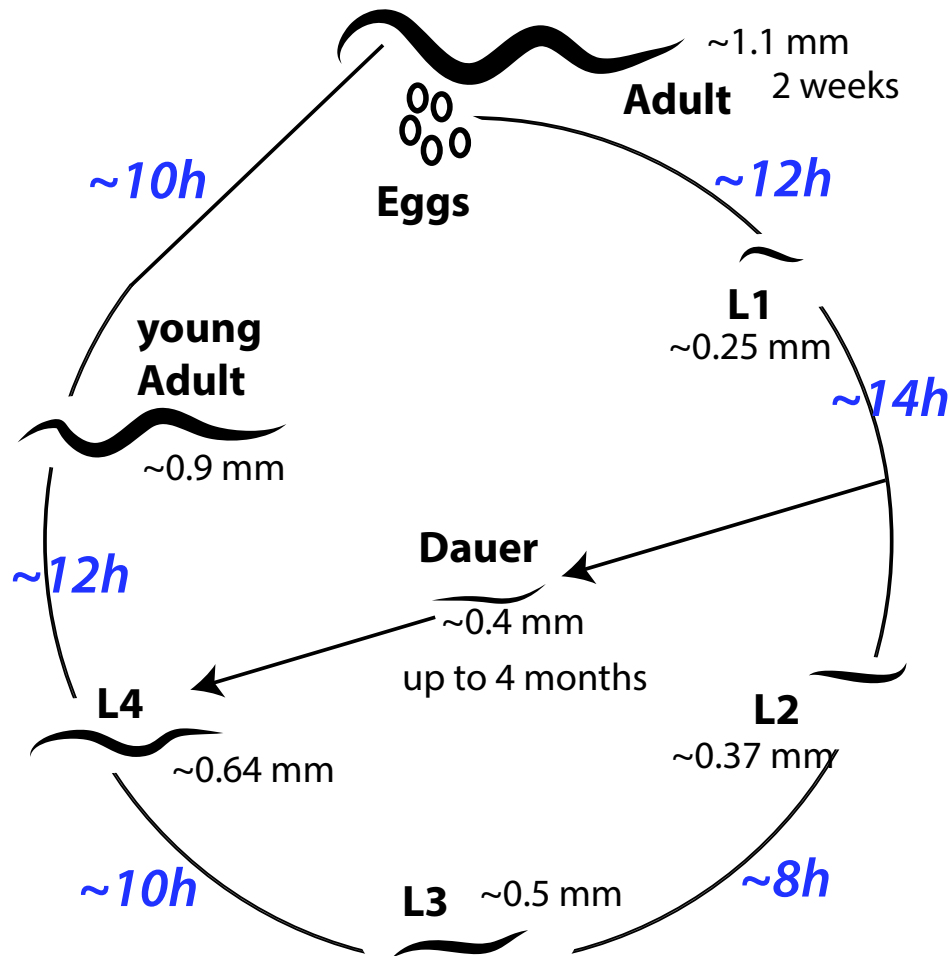


Figure I.3.: *C. elegans* lifecycle:

Different developmental stages of *C. elegans*, including their approximate timing at 20°C and the worm's length. Numbers are taken from Byerly *et al.*[78]

distinct dauer larval stage is an arrested stage which L2 larva may enter if environmental conditions become unfavourable [492]. Such conditions are high temperatures ($> 25^{\circ}\text{C}$), scarce availability of food, and high population density, sensed by the increased concentration of small molecules, the dauer pheromones [203]. Dauer larvae have altered metabolism [72] and reduced locomotion, and may survive up to several months and develop directly into L4 larvae once environmental conditions improve. Despite having a normal lifespan after escaping the dauer stage, it was shown that during the dauer stage, worms actually age [254] and can even die, *i.e.* are not able to recover to L4, if conditions do not allow lifecycle progression.

Not only the dauer stage makes *C. elegans* an interesting model for ageing and longevity, but also the presence of long-lived mutant strains. In 1988, the first mutation which increases lifespan was found; it affects the gene *age-1*, which encodes for a catalytic subunit of the PI3K orthologue [182]. In 1992, a mutation that doubles life-span by reducing insulin like signalling (ILS) was discovered by Kenyon and coworkers [310]. Since then, *C. elegans* has become one of the most important model organisms for ageing. Using large scale RNAi screens, many more genes have been discovered that when deleted or over-expressed extend lifespan [224, 227, 351, 439, 415]. These genes act over many different pathways. An extensive review of all those pathways, can be found in the introduction of Chapter V., in which the metabolic profiles of several life extending mutations are compared.

Most of these life-span enhancing mutants have been extensively studied from a genetic control [311] and transcriptomic point of view [377]. However, little research has been done on the role of metabolism itself. This is surprising, considering that in addition to genetic factors, there are also a number of small molecules and physicochemical conditions [456], that can significantly increase *C. elegans* lifespan. Those small molecules include valproic acid [155], resveratrol [34], spermidine [424], oxalacetate [626], glycolysis inhibitors [518], antidepressants [478], antidiabetics [80], or different carbohydrates [249], such as trehalose. Also, life expectancy is determined by the food source, such as by the choice of bacterial strains [382, 488] or whether they were alive or killed before feeding [601, 162].

Thus, analyses which focus on metabolic changes, such as metabolomics, may provide interesting insights.

I.3. Systems biology and metabolomics

Over the past century molecular biology and biochemistry have enabled scientists to understand many fundamental processes and principles, which when combined, make up an organism. Both disciplines usually work by reductionism. This means that they try to isolate and abstract a mechanism as much as possible, in order to understand its function. Therefore, those disciplines may be viewed as bottom-up approaches. Even though such simplified systems are an effective way of studying underlying mechanisms, they may fail

to capture integrated biological regulation. In any organism, even as simple as a single cell, those systems are all interconnected, regulating and impacting each other in a regulated but often non-linear manner [359, 91].

More recently, enabled by the rapid increase of available computing power, progress in database curation, modern chemical analysis and sequencing technology, science has begun to use top-down approaches to understand this complex interplay [621], and ultimately reach a more holistic picture of a biological system. The emerging field of systems biology aims to pool methods and knowledge from various classical disciplines of science, such as biology, chemistry, physics, mathematics and informatics, to build large network-based models, connecting the individual mechanisms. Sub-disciplines of systems biology are usually referred to as "-omics". [269, 328]

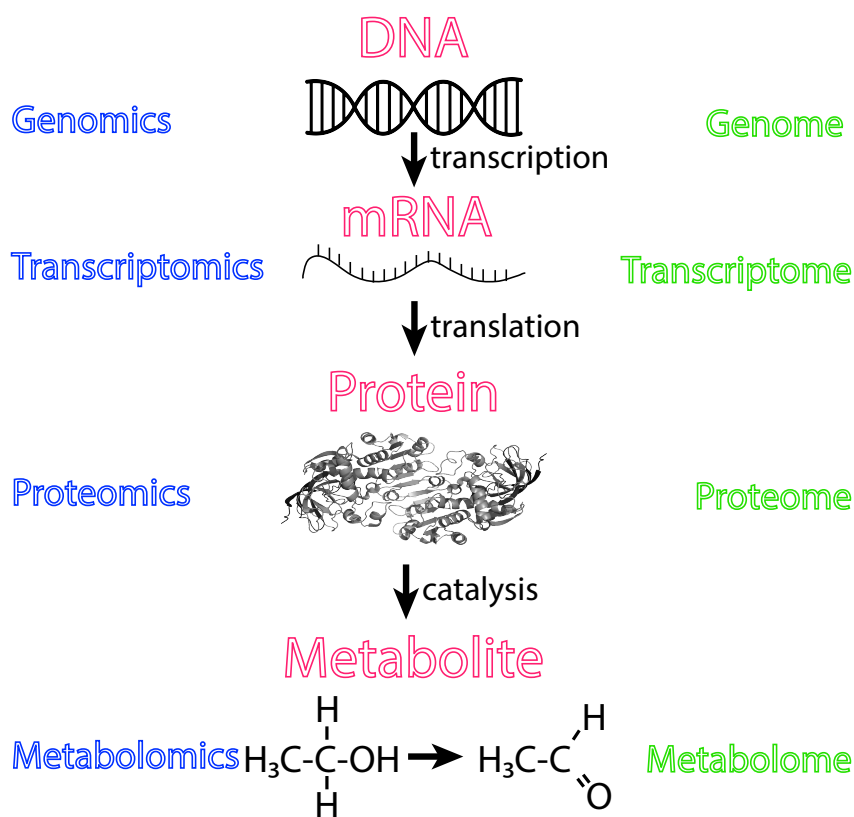


Figure I.4.: System biology's OMICS

Many different (or at least proposed) omics disciplines exist. The most common, however, are centred around Francis Crick's central dogma of molecular biology [119], as shown

in figure I.4. Genomics uses high throughput sequencing technology with bioinformatic annotation to determine the DNA sequence and proposed gene functions of an organism [272]. Transcriptomics aims at identifying and quantifying all expressed transcripts inside a system. Unlike genomics, transcriptomics at least partially reflects an organism's environment, as only transcripts currently being expressed are measured. Transcriptomics was typically performed by expression microarray analysis, but next generation sequencing based RNA-seq has become the gold standard [612]. Proteomics aims at determining and quantifying all expressed proteins [583]. Proteomics is even more specific than transcriptomics, as post-translational modifications, such as phosphorylation, ubiquitination or glycosylation, may also be assessed. Methods may range from gel electrophoresis with subsequent antibody detection, to sophisticated mass spectrometry based assays.

Metabolomics, the untargeted profiling of metabolites [122, 444, 171, 204], is the most downstream of system biology's omics approaches. It aims to quantify and identify low molecular weight compounds (e.g. below 1 kDa), such as organic acids, carbohydrates or lipids, inside a biological matrix. As metabolites are an endpoint of the biological system, metabolomics gives an integrated phenotype, reflecting the effects of environmental factors [204]. Thus, it provides an ideal tool to study complex phenomena like ageing, which are the result of genetic predisposition impacted by environmental conditions and various stimuli.

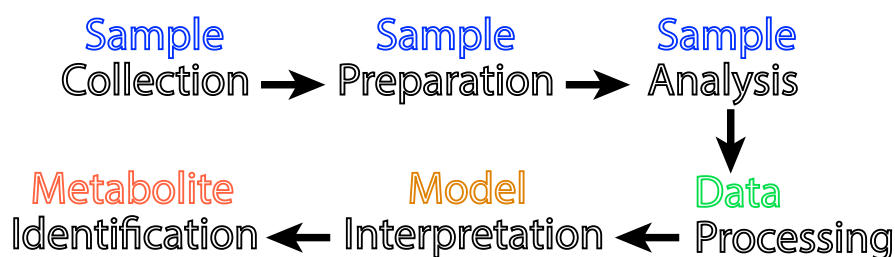


Figure I.5.: **Workflow of a typical LC-MS experiment**

Indeed, many different sample types have been analysed, using different analytical tools, modelling and analysis approaches in order to answer questions from very different fields of science [246, 553]. The typical workflow of a metabolomics experiment is outlined in figure I.5. It starts with sample collection, with the most common sample types being animal and plant tissues and biofluids, such as urine or blood plasma. Sample preparation may range from minimal intervention, to sophisticated extraction and derivatisation procedures. The

most commonly used analytical platforms are nuclear magnetic resonance spectroscopy (NMR) [442], and gas or liquid chromatography, coupled to mass spectrometry (GC-MS, LC-MS) [135, 614]. Data processing may range from quantification of a targeted set of metabolites up to complex dataset deconvolution and alignment procedures. Analysis may also range from classic biomarker discovery, using multivariate pattern recognition [579], time series analysis [477, 11, 329, 148], up to quantitative pathway based models [457, 279, 175, 167].

All elements of this workflow, from the basic science, over instrumentation up to experiment execution with subsequent processing and statistics are discussed in detail in the theory and methods chapter (VI.4), which was moved to the appendix.

The combination of advanced analytical approaches and sophisticated data processing and analysis tools is key to metabolomics experiments. Thus, analytical and computational method development are a main aspect of this thesis.

I.4. Metabolomics of ageing and *C. elegans*

Metabolomics has also been employed for ageing research [417]. It has been used to find biomarkers of ageing and longevity, for several model organisms, such as yeast [655], fruit flies [514], mice [256, 128, 625, 573] and even dogs [611], up to human cohort studies [213, 657]. A large proportion of all reported significant molecules are lipids, analysed by reversed phase chromatography.

C. elegans ageing has also been subject to metabolomic analyses. Fuchs *et al.* [183] characterized a "metabolic signature of long life in *Caenorhabditis elegans*" by applying a ^1H -NMR metabolomics approach to IIS longevity mutants and worms of different ages. They found branched chain amino acids (BCAA) were a key set of metabolites. Martin *et al.* [392] independently reported very similar findings, using the same or related mutant strains. An *et al.* used multidimensional NMR to also show the involvement of BCAA in sirtuin based life-span extension [15]. Butler *et al.* showed that long-lived mitochondrial mutants use a novel type of mitochondrial metabolism [77] with the involvement of alpha-ketoacids [75].

Most of these findings were published only after the start of this degree. Furthermore, the majority of long-lived mutant strains have not yet been examined by metabolomics. A key shortcoming of current published studies is that they focus on single pathways of lifespan extension, without comparing between them. Hence, analysis of additional long-lived mutant strains and making comparisons between them will be the main application for the methods developed in this thesis. This is described in a large scale longevity mutant screen in chapter V..

Next to ageing, other aspects of *C. elegans* biology have been examined using metabolomics. Those include cell growth [556], fat metabolism [23, 86, 85], heavy metal and pesticide toxicology [264, 288] and reactive oxygen species [660]¹ and acute anoxia [76] responses, respectively. Reinke *et al.*[488, 558] showed that the choice of *E. coli* strain affects the *C. elegans* baseline metabolome. Davies *et al.*[125] also showed that the mitosis inhibitor fluorodeoxyuridine (FUdR), which is often used to maintain synchronous populations of *C. elegans* during longevity-assays, heavily impacts the metabolic phenotype.

Blaise *et al.*[51] also developed a protocol for whole worm metabolite profiling by magic angle spinning NMR. However, apart from these last three studies, there has been surprisingly little method development, considering the rapid growth of *C. elegans* metabolomics research. This gap in the literature will be addressed by this thesis, in which analytical and computational methods for *C. elegans* metabolite analysis will be developed, with an emphasis on increasing metabolite coverage by the use of multiple analytical platforms. Many of the methods will also be applicable to other model organisms, and also the broader field of metabolomics.

I.5. Aims and structure of this thesis

The aim of this thesis is to address the lack of advanced methods for *C. elegans* metabolite analysis; with conjoint analytical and computational approaches in particular. Those methods will be used to understand the metabolic basis of factors that drive ageing and longevity in *C. elegans*.

¹This work was published in collaboration with Stephen Stuerzenbaum and Sukaina Zeitoun-Ghandour from Kings College, London; Metabolite profiling and data analysis was performed by Florian Geier.

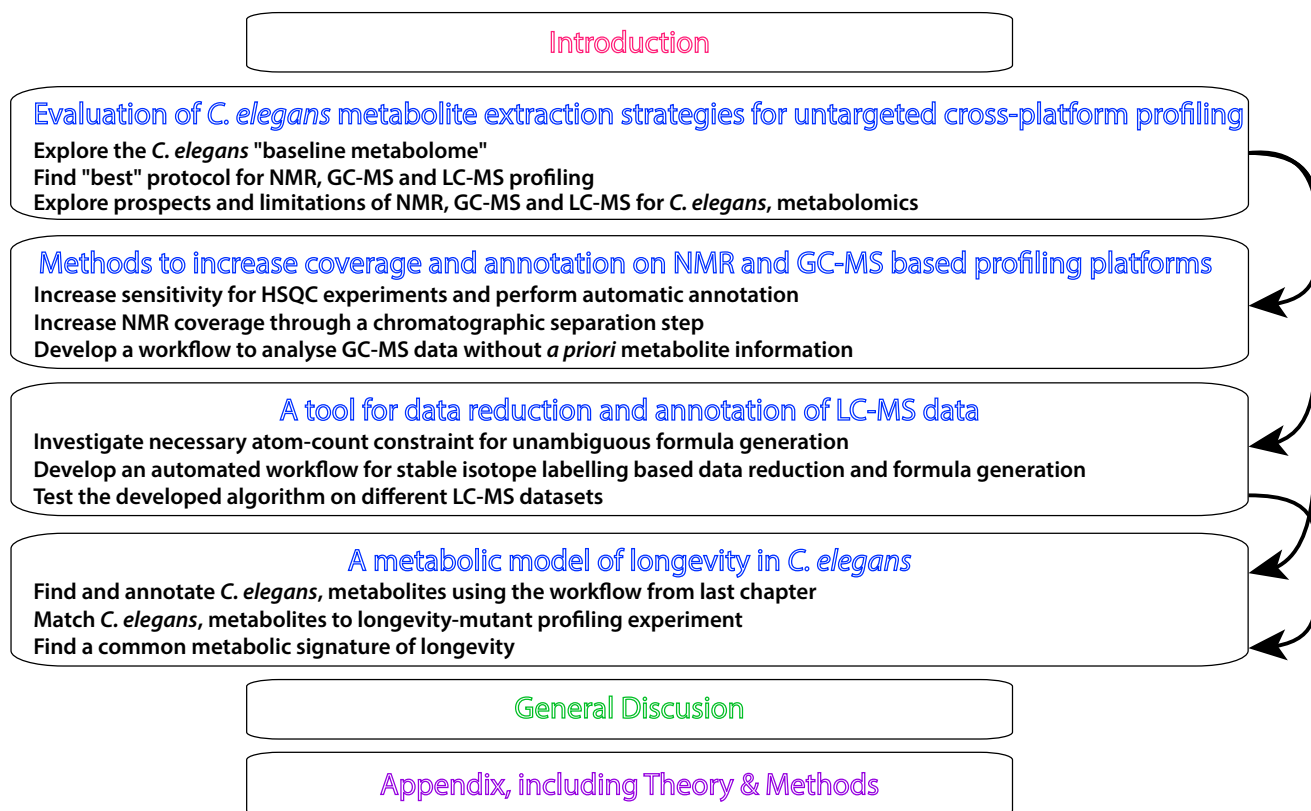


Figure I.6.: Overview of thesis structure

Figure I.6 outlines the general structures of this thesis and the specific objectives of the individual chapters.

The thesis has four results chapters, of which the first three are primarily concerned with method development, the last one, with the application of some of those methods.

In the first results chapter (II.), *C. elegans* metabolite extraction protocols will be benchmarked for subsequent GC-MS, NMR and LC-MS based untargeted profiling.

The second chapter (III.) is a compound chapter, containing three mini-projects, to increase metabolite coverage and annotation for NMR and GC-MS. The first project (III.2), uses ^{13}C to enhance the sensitivity of 2-D NMR spectra (HSQC) and perform automatic database annotation of those. In the second project (III.3), a LC-NMR ion-pairing strategy was used to separate the anionic constituents of the NMR visible *C. elegans* metabolome;

a NMR-silent, fully deuterated ion-pair reagent was evaluated. The third project (III.4) proposes an alternative to the standard GC-MS analysis workflow from chapter one. The workflow dynamically aligns and overlays the total ion chromatograms (TIC) of different sample groups for a quick, visual analysis. This method will be applied to the analysis dataset of mitochondrial mutants, which make *C. elegans* either long- or short-lived.

The third results chapter (IV.) is concerned with developing a method, which reduces LC-MS data to biological, non-redundant features only, which are subsequently annotated with their molecular composition. This method relies on the increase of molecular weight, when a metabolite incorporates stable isotopes. First, a simulation explores the feasibility of this approach, then protocols for triple (carbon, nitrogen, sulphur) stable isotope labelling of *E. coli* and *C. elegans* samples and a software algorithm to extract this information from LC-MS data, is developed. Finally the method is tested with *E. coli* samples, acquired by different chromatographic and electrospray modes.

In the final experimental chapter (V.), a panel of 24 different long-lived *C. elegans* mutant strains is analysed, to find a common metabolite pattern, associated with longevity. This is done in a "targeted untargeted" fashion. First, the stable isotope based data reduction and annotation method is used, to annotate a *C. elegans* extract. Then, the annotated biological features are integrated within the mutant dataset. The resulting data matrix is then subject to a variety of multivariate and univariate modelling approaches, also considering the annotated molecular compositions.

In a final discussion chapter (VI.), results will be summarised and put into the context of current issues faced in metabolomics and (*C. elegans*) ageing research. Also, two methods, which were already published as a proof of concept, and promise to solve key problems in *C. elegans* ageing research, will be presented.

In the appendix (A), a comprehensive discussion on the theory of the methodology used can be found. This chapter also gives an introduction to the algorithms, including their parameters, used for data-processing. Thus, it may serve as a general source of reference. Next to the supplementary tables and figures (B), the appendix contains wet-lab protocols, for procedures, which are applied frequently throughout the thesis (C).

II.. Evaluation of *C. elegans* metabolite extraction strategies for untargeted cross-platform profiling

Aims and Objectives:

- Explore the *C. elegans* "baseline metabolome"
- Find the "best" protocol for NMR, GC-MS and LC-MS profiling
- Explore prospects and limitations of NMR, GC-MS and LC-MS for *C. elegans* metabolomics

II.1. Introduction

The ultimate goal of metabolomics is to measure the concentration of each metabolite contained inside a matrix. Endogenous metabolites have a high diversity of physicochemical properties compared to other biomolecules, such as RNA or DNA. Thus, one sample simultaneously contains small, highly polar compounds from central carbon metabolism, amphiphilic amino acids, and high molecular weight, hydrophobic lipids. This has two major implications. Firstly, as discussed in A2.4, a combination of analytical methods is required to provide the most comprehensive coverage of metabolite classes. Secondly, sample preparation, including the choice of solvents, which themselves bias particular metabolite classes, is crucial. Whereas sample preparation for biofluids, such as urine [616] or blood [615, 141], is often minimal, with typical steps such as sample dilution or protein precipitation, preparation of tissue samples is more complex. Tissues must be disrupted to allow the metabolites to be accessed by the extraction solvents.

C. elegans are both tough and small, with an adult size of $1.2 \text{ mm} \times 60 \text{ }\mu\text{m}$, and an external collagenous cuticle. As a result, tissue extraction protocols [613] developed for eukaryotic cells or tissue samples from larger organisms are unlikely to be suitable for *C. elegans*, with tissue disruption being a particular concern. There are many extraction protocols in the metabolomics literature, but little has been done with regards to *C. elegans*. Bhaskaran *et al.*[47] showed that a Balch homogeniser delivers good results for protein and DNA extractions. However, their device operates on a "one sample per batch" basis and would therefore not be suitable for high throughput experiments. To find the best metabolite extraction method for *C. elegans*, this chapter benchmarks six different tissue disruption methods.

Following tissue disruption, metabolites must be solubilised by the extraction solvents. Equally important, the extraction solvents must effectively halt all metabolic activity, to enable a reproducible, accurate snapshot of the metabolome: aqueous extractions are generally not suitable, and may lead to significant degradation of many metabolites within seconds after disruption [364]. Thus, for most metabolomics applications a mixture of aqueous and miscible organic solvents is used. Most commonly, those are either monophasic or bi-phasic. Bi-phasic systems, however, lead to metabolite partitioning, such as separation between hydrophilic and hydrophobic compounds, between the two phases. [349, 394, 364, 367, 613].

II.2. Design of Experiment

This chapter aims to evaluate *C. elegans* extraction strategies for subsequent untargeted profiling using NMR, GC-MS and LC-MS, focusing on tissue disruption techniques, which, in contrast to solvent systems, have received little attention throughout the literature.

The success of a solvent - disruption pair was assessed by three criteria.

Similarity What is the common or unique metabolite coverage

Reproducibility How does within-replicate variation compare

Yield Which method extracts "the most" of a certain metabolite

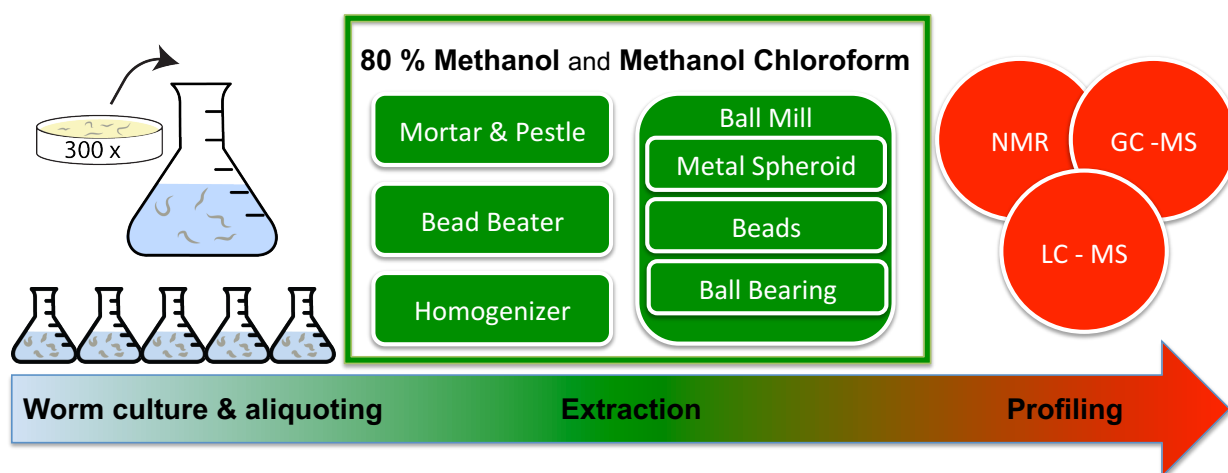


Figure II.1.: **Design of experiment:**

300 agar plates of *C. elegans* were pooled and aliquoted into 60 fractions for 5 replicates of six extraction methods and two solvent systems. After extraction with either 80 % methanol or methanol:chloroform:water, each extract was aliquoted for GC-MS, LC-MS and NMR in a ratio of 1:1:8.

The full-factorial design of the experiment is outlined in figure II.1. 300 plates of mixed N2 wild type worms were pooled and aliquoted, and then extracted using 2 solvent systems and 6 tissue disruption methods, with 5 replicates each.

A mono-phasic 80 % methanol-water (M80) mixture and a classic bi-phasic Bligh & Dyer [52] methanol-chloroform-water (MCW) approach were selected as extraction solvents. Both of those have successfully been used to extract metabolites from *C. elegans*[556, 183, 125, 23, 86], but their performance has not been compared yet. For the

MCW extraction, only the polar aqueous fraction was tested, as specific protocols for *C. elegans* lipid analysis already exist [396]. For tissue disintegration six methods were compared, with four different pieces of equipments, one tested with three different milling agents:

(1) Manual grinding with mortar and pestle, (2) bead-beating in a planetary bead-beater, as widely used for DNA/RNA/Protein extractions, (3) homogenisation by a blender, (4) a grinding shape optimised metal spheroid inside a reciprocating ball mill, (5) silicon beads inside a reciprocating ball mill, and (6) grinding at cryogenic temperatures using a metal ball bearing inside vessels, operated inside a reciprocating ball mill.

All samples were then analysed by GC-MS, LC-MS and NMR.

II.3. Materials and Methods

II.3.1. Sampling and extraction

300 plates of N2 Bristol wild type worms were grown to a mixed age population. Each plate was washed with 5 ml of sterile M9 buffer and the worm suspension pooled inside a large beaker. After the worms had settled, the volume was reduced to 25 ml and dispensed into 60 (+ 2 backup) 0.5 ml aliquots under constant shaking to ensure homogeneity. Aliquots were snap frozen immediately and stored at -80°C, until extraction.

As extraction solvents a mono-phasic 80 % methanol (M80) and a bi-phasic Bligh & Dyer [52] 2:1:1 methanol:chloroform:water (MCW) extraction were chosen. Six tissue disruption methods were compared: (1) manual grinding, using a liquid nitrogen chilled ($\varnothing = 10$ cm) mortar and pestle, (2) homogenisation by an electric homogeniser, (3) bead-beating, and finally beating in a reciprocating ball mill (Qiagen Tissue-Lyser (Qiagen, Crawley, UK)) used with (4) metal spheroids, (5) SiO_2 beads, or (6) metal ball bearings at cryogenic temperatures. The exact procedures on how these were operated are outlined below.

In all cases a two-step extraction was used to improve recovery: first, a direct extraction of the worm pellet, and second, another aliquot of solvent was used for rinsing (mortar and pestle, homogenizer, ball bearing) or re-extraction of the pellet (bead beater, spheroids, glass beads). For the M80 extraction, 2 ml of methanol were added to the 0.4 ml aqueous worm pellet, so that a 80 % methanol composition was reached. For the second step (rinsing/re-extraction) 1 ml of 80 % methanol was used. In the first step of the MCW extraction, 1.2 ml of 2:1 v/v methanol:chloroform were added to the 0.4 ml pellet, to give a 2:1:1 methanol:chloroform: water mixture. For the second step, 0.8 ml of 1:1 chloroform:water were used.

Generally, extraction times and intensities were chosen based on manufacturers' recommendations, existing protocols from other sample types, and also, that with the amount of possible cooling, the sample temperature never exceeded 0°C. Exact procedures are as follows.

The...

Mortar & pestle were cleaned with ultra-pure water, following a 0.1 M HNO_3 acid wash between each sample. Using liquid nitrogen, they were pre-chilled before grinding

the worm pellet to a fine powder (*ca.* 10 min). Extraction solvents were directly added to the mortar and recovered after dissolving the frozen powder, using a glass pipette.

Homogeniser was a T25 ultrathurax (Janke & Kunkel, Staufen, Germany), operated for 3 min at 24×10^3 rpm inside a 10 ml glass vial. To ensure that the homogeniser's blades were immersed inside the extraction solvents at all times, the solvent volumes had to be increased. This also required a change in solvent ratios to yield the same final solvent composition as in the other methods, when considering the 0.4 ml aqueous worm pellet. For the 80% methanol extraction, 3.6 ml 90% were used for the first and 2 ml 80% MeOH for the second step. For the MCW extraction, volumes were increased to 3.6 ml of 2:1:0.4 methanol:chloroform:water and 2 ml of 1:1 chloroform:water in the first and second step, respectively.

Bead-beater was a FastPrep 12 (MP Biomedicals, Cambridge, UK), operated at 6.5×10^3 rpm for 2 times 30 s and a 60 s cooling step in-between. The frozen pellet was split into two 2 ml tubes (StarLabs, Milton Keynes, UK), after filling those with *ca.* 100 μ l of 0.1 mm glass beads(VWR). The cold extraction solvents were added and the extraction started before the pellet could thaw.

Metal oblate spheroids of *ca.* 3 mm diameter have superior grinding performance compared to spherical balls, according to the manufacturer (Retsch, Leeds, UK). One spheroid was placed inside a 2 ml microcentrifuge tube, before the worm pellet was added. Again, the frozen pellet was split into two tubes before addition of the cold solvents. The ball mill was operated at 25 Hz for eight min.

Glass beads were identical to those, used inside the bead-beater. The extraction procedure was the same as for the metal spheroids, just with the beads instead of the spheroids.

Ball bearing with a diameter of 7 mm were put inside 5 ml stainless steel vessels (Retsch) (one per vessel), before submerging the vessel into liquid nitrogen for 20 min. The sample was loaded and the ball mill operated as for the metal spheroids and glass beads. The ground powder was then extracted directly.

After extraction, all extracts were combined (where split) and centrifuged (5 min, at 16×10^3 g, 4°C) to precipitate residual protein. The supernatants were pooled inside a centrifuge

tube and - if necessary, topped up to a defined volume (2 ml and 6 ml for the homogeniser extracts) to allow precise aliquoting. After aliquoting for GC-MS, LC-MS, and NMR in a ratio of 1:1:8, the samples were dried overnight in a vacuum sample concentrator (Eppendorf, Hamburg, Germany) at room temperature. All samples were kept at -80°C until analysis. Aliquots designated for GC-MS analysis were dried directly inside silanised GC-MS glass vials (Agilent).

Only the aqueous fractions of the MCW extraction were analysed in this study, as targeted lipidomic protocols for *C. elegans* are available [396].

II.3.2. Analytics

II.3.2.1. NMR

Sample preparation, was performed according to the standard procedures, described in the appendix B3.

Spectra were acquired on a Bruker Avance III spectrometer with a 14.1 T magnet (Bruker Biospin; Rheinstetten, Germany), operating at 600.13 MHz ^1H observation frequency, and using a cryogenically-cooled inverse 5 mm probe; the spectrometer was controlled using Bruker TopSpin 2.1.

Acquisitions were based on a standard ^1H NOESY [RD-90°- t_1 - 90- t_m -90°-acquire] sequence. The recycle delay (RD) was 3.5 s, with a mixing time (t_m) of 0.1 and acquisition time of 1.3 s. The 90° was determined for each experiment but usually was between 8 – 14 μs . 128 free induction decay (FID) transients were collected and summed over a 20 ppm sweep width from -5 to 15 ppm, with a time domain resolution of 32k data points. The water signal was suppressed by pre-saturation during the RD.

For peak assignment, coupling constant resolved (J-Res), ^1H - ^1H correlation spectroscopy (H,H-COSY) and heteronuclear ^1H - ^{13}C single quantum bond correlation (HSQC) spectra were acquired.

Furthermore, 32 assigned metabolites were quantified using a computer-assisted manual fitting approach [617]. To this end, the commercial software package Chenomx NMR Suite

6 (Chenomx, Alberta Canada) was used as described in detail by Tredwell *et al.*[578].

II.3.2.2. GC-MS

Derivatisation, acquisition and processing were performed according to the standard procedures described in chapter B4 and A2.1.3. To monitor runtime and machine variation, quality control samples (QC) [511] were injected every ten samples and five consecutive times in the middle of the run, respectively. Those QC samples were aliquots from the QC samples prepared for LC-MS analysis (See next section).

II.3.2.3. LC-MS

Dried extracts were solubilised by 120 μ l of water by using an ultrasonic bath for 5 min and vortexing. After centrifugation (5 min, at 16×10^3 g, 4°C), 100 μ l of the supernatant was transferred into a 96 well autosampler plate. The remaining volumes were pooled, centrifuged and used as quality control (QC) sample. 10 μ l were injected onto a 2.1×100 mm Acquity BEH C18 column (1.7 μ m) (Waters Corporation, Milford, USA) by a Waters Acquity ultra-pressure liquid-chromatography (UPLC) system (Waters Ltd. Elstree, UK). Analytes were eluted on a linear 25 min gradient from 100% water (+ 0.1 % formic acid) to 100% acetonitrile (+ 0.1 % formic acid) with a flow rate of 0.5 ml per min and a column temperature of 40°C.

The LC was coupled online to a LCT Premier mass spectrometer (Waters MS Technologies, Ltd., Manchester, UK). Spectra were acquired in negative electrospray ionisation mode (ESI-), with a capillary and sampling cone voltage of 2.4 kV and 35 V, respectively. Desolvation temperature was set to 350°C, source temperature to 120°C and desolvation gas flow (N_2) to 900 litres per hour. The instrument was operated in v-mode optics, for maximum sensitivity. Data were acquired at a scanning frequency of 10 Hz, with a 10 ms inter-scan delay over a scan range of 50 - 1000 u in centroid mode. For high mass accuracy, a 200 pg/ μ l leucine-enkephaline ($m/z = 556.2771$) was infused as lock-spray at 3 μ l per min and 3 lock-mass scans averaged every 15 s.

The whole LC-MS system was controlled using Water's MassLynx software suite.

In the beginning of the run, 5 QC samples were injected to condition the column. All samples were run in randomised order and QC samples were injected every 10 samples to assess runtime variation. In the middle of the run, 5 consecutive QC injections were made to quantify pure machine variation.

After acquisition, raw data were converted into netCDF using MassLynx. The free software package xcms [531] was used to detect metabolite features and align them across data files. For peak detection the continuous wavelet transform based algorithm "centWave" [566] was used with recommended UPLC parameters (peak width between 3 and 15 s and maximum retention time deviation of ± 2 s). During the lock-mass scans, no signal is recorded for the sample, resulting in gaps throughout the acquisition. Therefore, an algorithm developed by Benton *et al.* [43] was used to "stitch" these gaps in order to avoid distortion of the reproducibility and yield metrics.

After obtaining the final sample $\times$ feature intensity matrix, all further statistical analysis was carried out in the statistical software suite R.

II.4. Results

II.4.1. Contribution of bacterial biomass to total signal

As *C. elegans* feeds on an *E. coli* lawn, there might be a bacterial contribution to the overall signal. Standard *C. elegans* cleaning procedures such as sucrose gradient centrifugation or even simple consecutive buffer washes could prevent this [181, 337, 159] - but will also alter worm physiology and thus impact the metabolic profile. The severity of washing-stress may well depend on the experiment and would therefore need to be assessed in each case. To get around this, the validity of the sampling protocol was tested prior to the extraction evaluation. This was done by measuring the amount of possible "bacterial contamination" of an unwashed *C. elegans* pellet. After the harvested worms had settled to form a pellet (ca. 5 min), 0.4 ml of buffer - directly above the pellet were sampled and processed in the same way as the pellet itself (bead-beater extraction as described in II.3.1). Thus, the pellet would at most have as many suspended bacteria as the supernatant. ^1H spectroscopy was used to quantitatively compare the signal of the supernatant to the pellet extract signal. Results displayed in figure II.2 confirm that the bacterial contribution to metabolite concentrations is minimal.

II.4.2. Assigned metabolites

Before differences between the different extraction methods were assessed, an attempt was made to assign as many metabolites as possible. Table C1.1 shows a list of metabolites assigned by NMR and GC-MS, with their corresponding PubChem identifier and functional class.

For the 32 NMR based compound assignments, 1D and 2D (^1H , ^{13}C -HSQC, ^1H , ^1H -COSY) in-house recorded spectra, as well as spectra from the the Biological Magnetic Resonance Data Bank (BMRB) [585] and the Human Metabolome Database (HMDB) [630] were used. GC-MS metabolites were initially assigned by AMDIS and the "Fiehn Library". In cases where the automatic retention time and fragmentation pattern based assignment was not unambiguous, the Golm database (GMD) [335] was also consulted. Authentic standards for most of the database-assigned metabolites were run for final confirmation. Some metabolites (partially) decay into other compounds during derivatisation or analysis - which is usually not documented in databases. However, this must be considered for

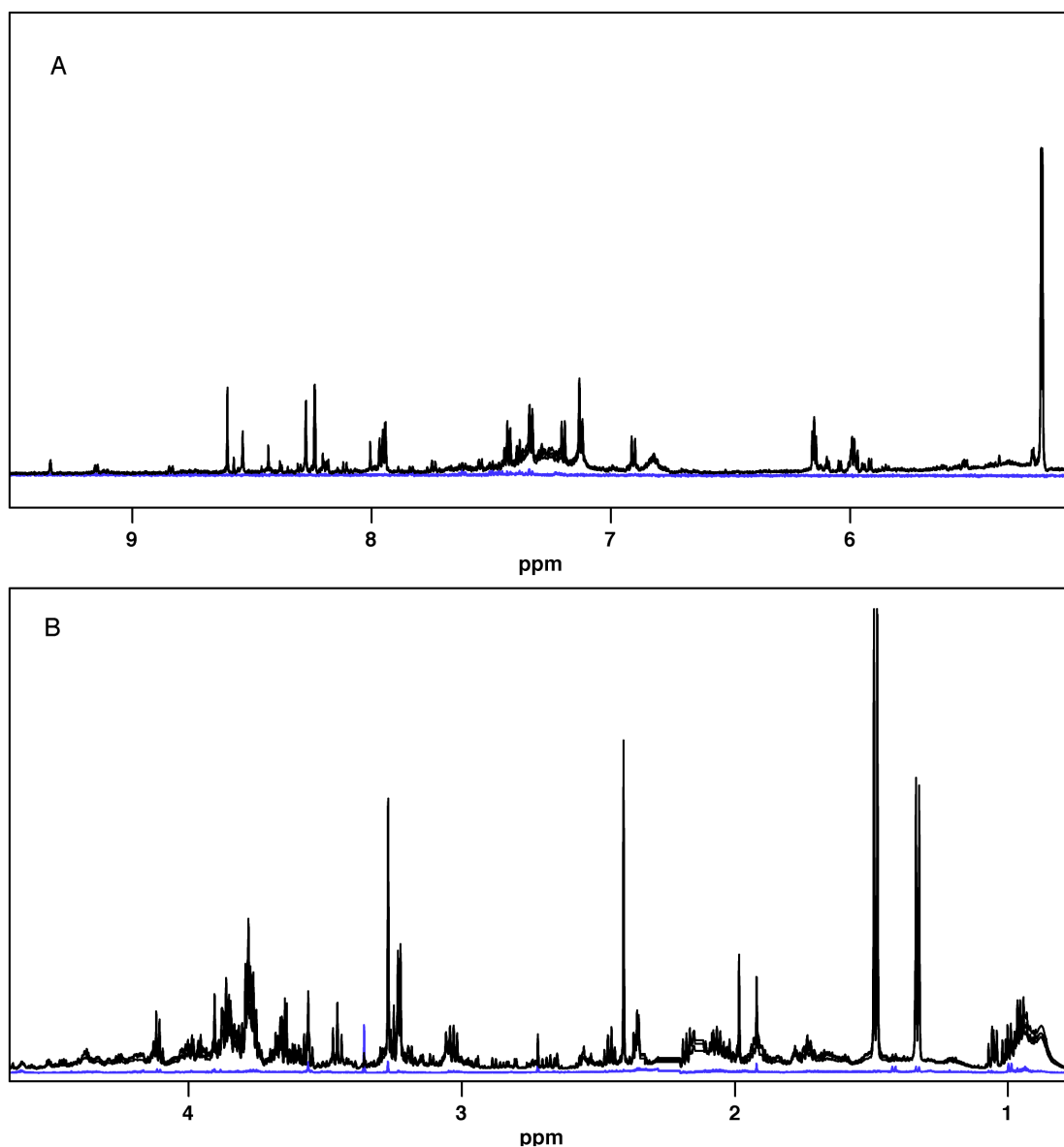


Figure II.2.: **Bacterial "contamination" of *C. elegans* pellet:**

Contribution of bacterial biomass to a *C. elegans* extract spectrum, as determined by 600 MHz ^1H spectroscopy. An equal volume of supernatant above an unwashed worm-pellet (blue) was sampled and compared to the pellet itself (black); $n = 3$ for both. The aromatic region (A) is magnified by a factor of 5, relative to the aliphatic (B).

Name	PubChem Key	GC-MS	NMR	Class
Betaine	247		x	Biogenic amine
Choline	305		x	Biogenic amine
O-phosphocolamine	1015	x		Biogenic amine
Phosphocholine	1014		x	Biogenic amine
Putrescine	1045	x		Biogenic amine
3-aminoisobutyric acid	64956	x		Amino acid
Alanine	602		x	Amino acid
Arginine	232	x*	x	Amino acid
Asparagine	236		x	Amino acid
Aspartate	424		x	Amino acid
Glutamate	611	x*	x	Amino acid
Glutamine	738		x	Amino acid
Glycine	750	x	x	Amino acid
Isoleucine	791	x	x	Amino acid
Leucine	857		x	Amino acid
Lysine	866		x	Amino acid
Methionine	876	x		Amino acid
Phenylalanine	994	x	x	Amino acid
Proline	614	x		Amino acid
Serine	617	x	x	Amino acid
Threonine	205	x		Amino acid
Threonine	205		x	Amino acid
Tyrosine	1153		x	Amino acid
Valine	1182	x	x	Amino acid
Beta - alanine	239	x		Amino acid Derivative/Precursor
Cystathionine	834		x	Amino acid Derivative/Precursor
D-Ala-D-Ala	601	x		Amino acid Derivative/Precursor
Homoserine	779	x		Amino acid Derivative/Precursor
Hydroxy-proline	5810	x		Amino acid Derivative/Precursor
Ornithine	389	x		Amino acid Derivative/Precursor
D-(+)-Trehalose	1143	x	x	Carbohydrate
Glucose	5793		x	Carbohydrate
3-Glycerophosphate	754	x		Carbohydrate Derivative/Precursor
3-Phosphoglycerate	724	x		Carbohydrate Derivative/Precursor
D-glucose-6-phosphate	5958	x		Carbohydrate Derivative/Precursor
Gluconic acid	604	x		Carbohydrate Derivative/Precursor
Glycerol	753	x	x	Carbohydrate Derivative/Precursor
myo-Inositol	892	x		Carbohydrate Derivative/Precursor
sn-Glycero-3-phosphocholine	71920		x	Carbohydrate Derivative/Precursor
Sucrose	5988	x		Carbohydrate Derivative/Precursor
Acetate	175		x	Carboxylic acid
Citramalic acid	1081	x		Carboxylic acid
Citric acid	311	x		Carboxylic acid
Formate	283		x	Carboxylic acid
Fumarate	723	x	x	Carboxylic acid
Lactate	612	x	x	Carboxylic acid
Malate	525	x		Carboxylic acid
Phosphoenolpyruvic acid	1005	x		Carboxylic acid
Phosphoglycolic acid	529	x		Carboxylic acid
Propionate	1032		x	Carboxylic acid
Succinate	1110	x	x	Carboxylic acid
Heptadecanoic acid	10456	x		Fatty acid
Linoleic acid	3931	x		Fatty acid
Palmitic acid	985	x		Fatty acid
Stearic acid	5281	x		Fatty acid
Adenosine	191	x		Nucleoside
Guanosine	765	x		Nucleoside
Xanthosine	1189	x		Nucleoside
Adenine	190	x		Nucleoside base
Thymine	1135	x		Nucleoside base
Uracil	1174	x		Nucleoside base
Xanthine	1188	x		Nucleoside base
AMP	224	x	x	Nucleotide
ATP	238		x	Nucleotide
GTP	762		x	Nucleotide
Inosine 5-monophosphate	8582	x		Nucleotide
UMP	6030	x		Nucleotide
Pyrophosphate	1023	x		Phosphate
Nicotinic acid	938	x		Vitamin or derivative
Pantothenic acid	988	x		Vitamin or derivative
Cholesterol	304	x		Steroid
Methoxytryptamine	1833	x		Biogenic amine
Phenylethanolamine	1000	x		Biogenic amine

Table II.1.: Putative *C. elegans* metabolites detected by GC-MS and NMR:

Pubchem identifier and compound class is also given. Compounds noted with x^* show an instability during derivatisation and are therefore only of limited interpretational value.

interpretation. Such incompatibilities are also documented in table C1.1.

For LC-MS, 4905 features were detected in the 80% methanol and 3595 features in the methanol-chloroform-water extracts, totalling 5141 across all extraction methods and solvents. Due to these high numbers and the stringent criteria the metabolomics community has set for LC-MS assignments [551, 428], no attempt was made to assign any of the features. However, the R-based CAMERA package [338] is designed to estimate the number of "real metabolites", by annotating redundant features, such as fragments, adducts and isotopes. Out of the 5141 features, CAMERA assigned 760 as isotopes and 1667 as adducts, fragments or multimers. This leaves 2713 features (53 %) as metabolite candidates.

II.4.3. Distribution of LC-MS features

As no LC-MS features were assigned, methods were compared with regards to their number of features only. Figure II.3a shows that between 50 % and 70 % of all LC-MS peaks could be detected in either none (0/0) or all (5/5) replicates of each individual method. Remarkably, features detected in only one, two, three or four replicates of a method often had a coefficient of variation similar to peaks detected in all (five) replicates of a method. However, when inspecting the corresponding extracted ion chromatograms, in most of these cases a feature was present, but failed to meet the minimum default peak detection criteria of the xcms "centWave" peak picker ($\text{intensity} \geq 100$, $\text{SN} \geq 10$). Low intensity metabolite features often deteriorate overall reproducibility of LC-MS data [197, 512, 199, 572]. For this reason, classic analytical chemistry sets the limit of quantification (LOQ) to a multiple of the limit of detection (LOD), which itself is a certain number of standard deviations (*e.g.* three) above zero. As a research technique, metabolomics does not necessarily require such arbitrary limits. However, defining thresholds to ensure high data quality is still perfectly justifiable. Here, for all further analyses, it was decided to include only features detected in all five replicates of at least one method (5/5). This is a conservative approach, and means that while there will undoubtedly be some false negatives, we can be confident that the between-method comparisons are based on robust data. Even with this stringent criterion, 1000 to 1366 features per disruption method could be detected (figure II.3b).

Figure II.3b displays the number of shared, reproducible features between any two methods. In terms of common reproducible features, the mortar and pestle, spheroid, beads

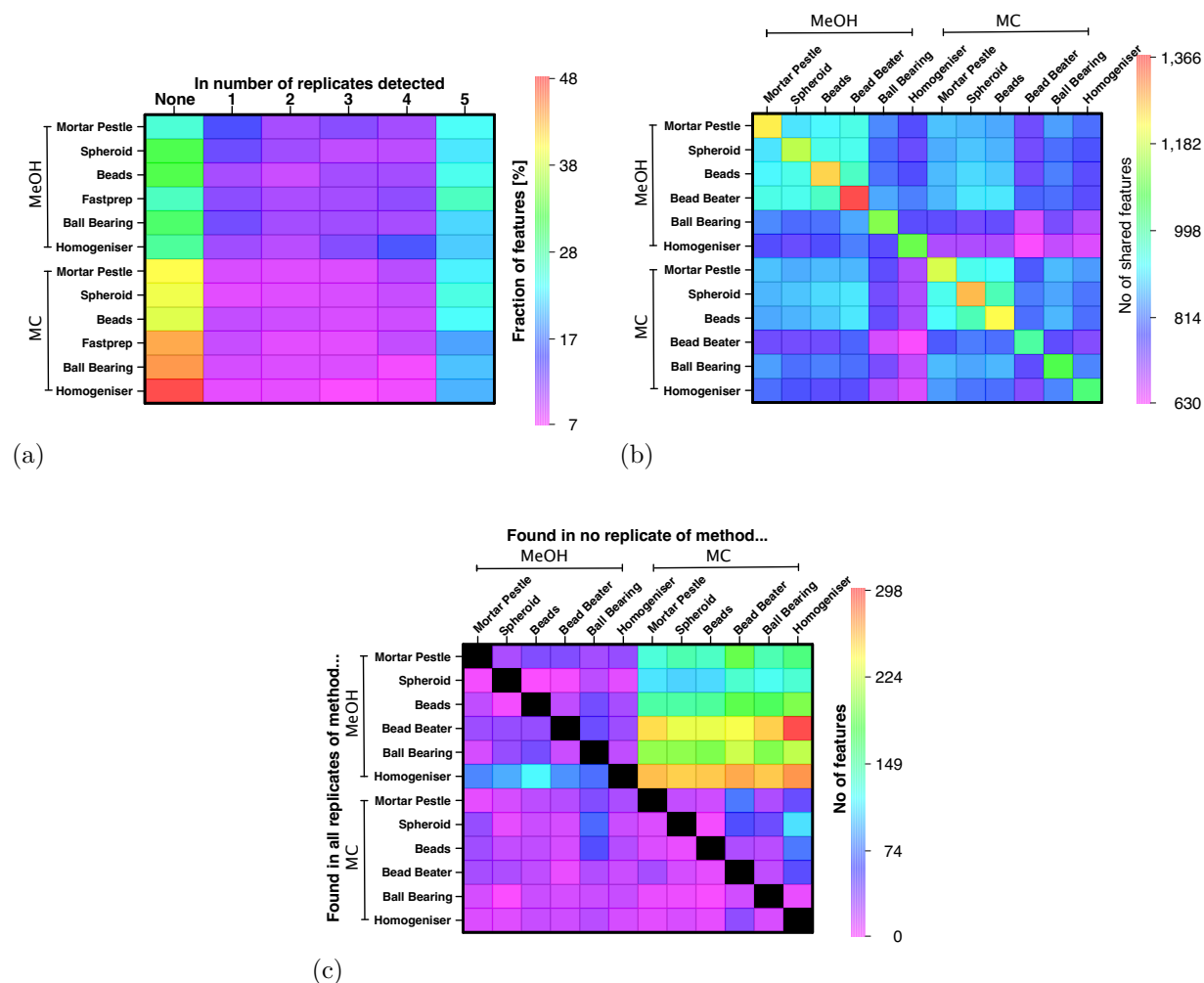


Figure II.3.: **Distribution of LC-MS metabolite features:**

II.3a: Detection of metabolic features across a method's replicates: The colour scale indicates the percentage of features, detected in a certain number of replicates of a method. For most extraction methods, most peaks are detected either in no replicate (0/5), or all of the replicates (5/5).

II.3b: The colour scale indicates the number of reproducible (5 out of 5 replicates) features, shared by two methods; with the diagonal representing the number of reproducible features of a method itself.

II.3c: The colour scale indicates the overlap between reproducible (5/5) and not detected features (0/5) between all methods, *i.e.* the number of features that are present in all replicates of one method, but not in any replicate of another method.

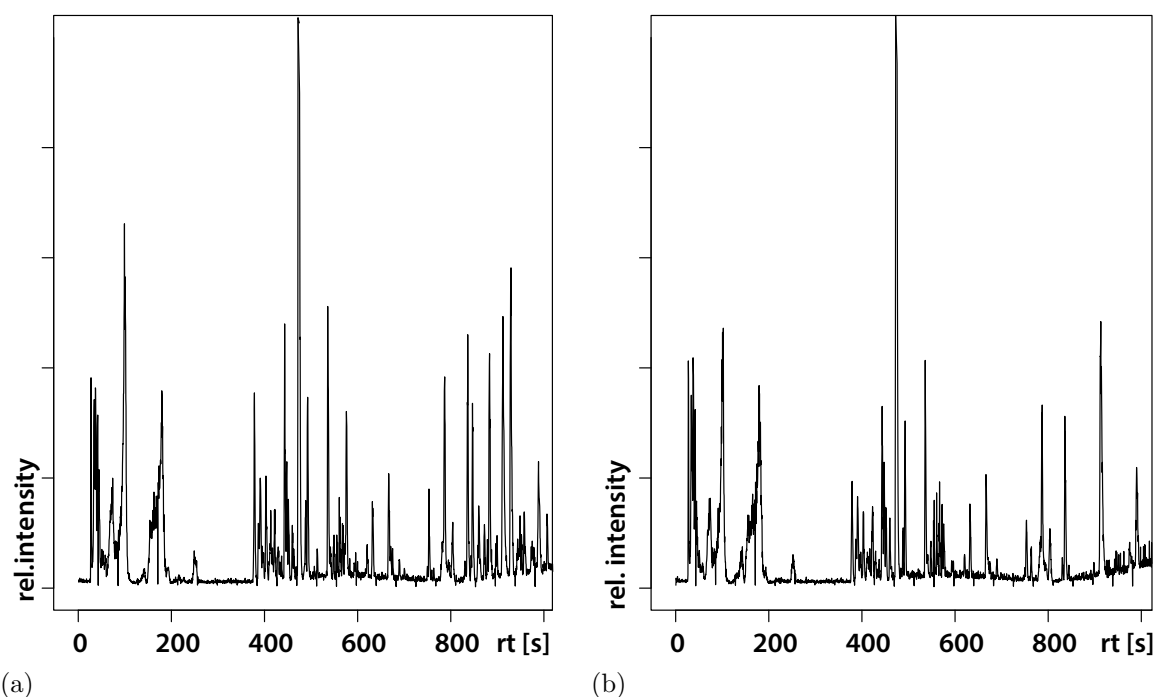


Figure II.4.: **LC-MS BIP chromatograms of M80 and MCW extraction:**

LC-MS base ion peak chromatograms for a representative M80 (figure II.4a) and MCW (figure II.4b) extraction. Whereas the chromatogram appears to be identical up to a retention time of 700 s, between 800 to 1000 fewer features are visible in the MCW sample. The periodic vertical lines are missing scans due to lock mass scans, which were corrected by using the gap filling algorithm from Benton *et al.*[43].

(TissueLyser) and bead beater (Fastprep) extraction methods showed the highest overlap within and across solvent systems, with the exception of the bead beater in conjunction with MCW. The ball-bearing and homogeniser extractions, however, were relative outliers, with the smallest overlap with any of the other methods. It is also interesting to ask which features were exclusive to a single method, *i.e.* the intersection of features detected in all (5/5) and no (0/5) replicates of a method. This comparison, displayed in figure II.3c, clearly shows that there are very few features unique to specific tissue disruption methods, but there are clear differences between the two solvents: M80 extraction resulted in about 100 to 300 additional reproducible features across all methods compared to MCW extraction. These are from lipids that are solubilised by methanol, but partition into the non-polar chloroform phase during MCW extraction. This is clearly demonstrated by the LC chromatograms, displayed in figure II.4.

II.4.4. Multivariate clustering and similarity

After a first exploration of the *C. elegans* baseline metabolome, it must be assessed how the different extraction methods compare. Before this is done in detail with respect to reproducibility and yield, multivariate clustering can be used to visualise broad similarities and differences between the methods. Scores of the principal component analysis of the log-scaled, mean centred data are displayed in figure II.5.

For both GC-MS (figure II.5a) and LC-MS (figure II.5b), runtime and (repeated injection) machine variance quality control samples clustered tightly. This indicates a robust analytical performance that allows further quantitative analysis of the data [511, 199, 197, 198, 44]. It is immediately apparent that for both mass spectrometric methods, samples separated with regards to solvent, but not tissue disruption method. In contrast, for the NMR data, neither the PCA of the aligned full resolution spectra (not shown), nor the PCA of the 32 fitted metabolite concentrations showed a clear solvent effect. However, the homogeniser used in conjunction with methanol-chloroform-water was clearly distinct from the rest of the other disruption/solvent pairs. Indeed, if the raw spectra of the homogeniser MCW method are compared to the other methods, some metabolites, *e.g.* trehalose, are completely missing, and others, like acetate, are highly increased. The homogeniser method was different from the other methods in two aspects. It used a larger solvent volume and the disruption process of the pellet did not start at a frozen state. As care was taken to reach the same final solvent ratio as the other methods, the most likely explanation is that these samples thawed slightly before extraction. Indeed, Liebeke *et al.*[364] reported that sample degradation may lead to drastic changes in metabolite concentrations within seconds, if enzymatic activity cannot be quenched.

II.4.5. Reproducibility

The multivariate analyses above already gave a first estimation of replicate to replicate variability of different methods. However, classically, the reproducibility of an analytical method is assessed using univariate methods.

Reproducibility itself is divided into accuracy and precision (see appendix figure A2.1). Precision is the deviation of results from their arithmetic mean, usually expressed as stan-

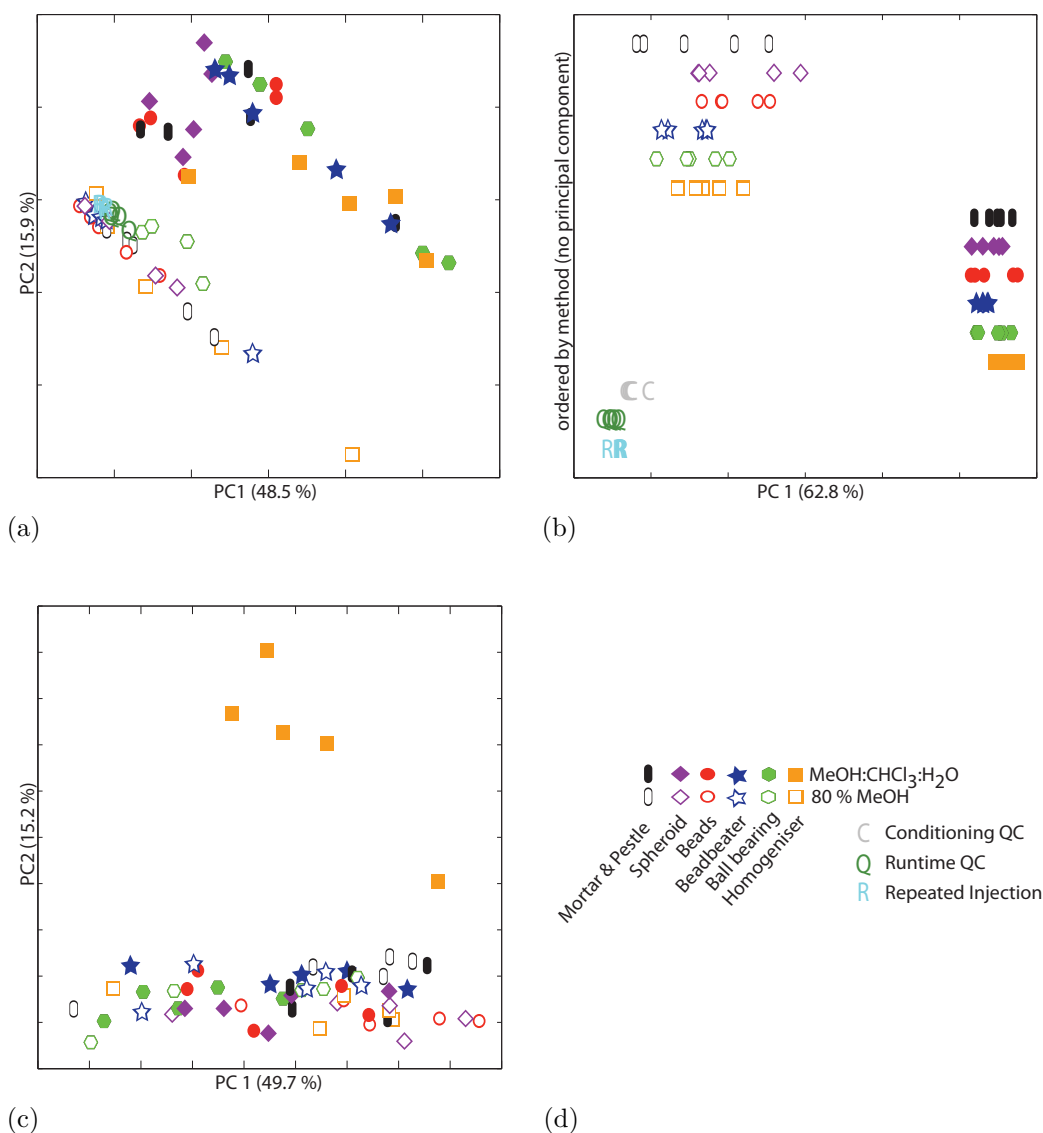


Figure II.5.: **Principal component analysis of extracted metabolite features:** PCA scores plots of the metabolites or metabolite features for GC-MS (II.5a), LC-MS (II.5b) and NMR (II.5c), including the R^2 for each principal component. For the LC-MS data, only one principal component was fitted, as PC 2 only explained 6 % of the variance, and did not provide any interpretational value.

dard deviation or coefficient of variation of the measured values. Accuracy, however, is the deviation of the measured from its "true" value. Whereas accuracy is relatively easy to assess for synthetic mixtures, such as drugs in pharmaceutical preparations, it is more difficult or even impossible for complex biological mixtures as the "true" concentration is unknown. For this reason, only precision, expressed as coefficient of variation (CV), was used to analyse reproducibility.

Figure II.6 displays the CV box plots of each method for the metabolites quantified by GC-MS (figure II.6a) and the ones fitted in the NMR spectra (figure II.6b). For LC-MS, only features detected ($SN \geq 10$) in all (5/5) replicates were used for this comparison. Hence, each method ends up with a different number of features. This could bias interpretation of the box plots, as illustrated by the following hypothetical example. Imagine two methods were to yield the same features with same CV, but one of them had additional features, with an average CV, worse than the average CV of the shared subset. In this case, this method would have its CV distribution skewed towards a higher value and thus be classified as less reproducible. However, it is as reproducible for the shared subset, but only has a higher coverage (for which it would be punished). For this reason, a ranked CV representation was used for method comparison in preference to boxplots, as displayed in figure II.6c.

Both GC-MS and LC-MS had a median repeated injection (RI) and quality control sample (QC) variability of 5 % and 10 %, respectively. No published data on the CV distribution of repeated injection (instrument) variation could be found for GC-MS. The mean technical reproducibility of GC-MS QC samples, including derivatisation induced variance, was shown to have an average CV of 8 % [171]. The reported median CV of instrument variation for LC-MS, however, is higher [661, 118]. This is most likely due to the fact that I employed a stringent minimum quantification criterion (feature detected in 5 out of 5 replicates). The median CV for the QC agrees well with previously published results for both GC-MS and LC-MS [531, 197, 394]. The CVs of the LC column conditioning QCs were nearly identical to the runtime QC samples. This could be taken to imply that our matrix may not require column conditioning, but it is still recommended to ensure stable chromatography and reproducible retention times. It is also noteworthy that the QC samples yielded 50 % more features on average than any of the methods. This could either be due to concentration effects in the QC sample, or that each method extracted a slightly

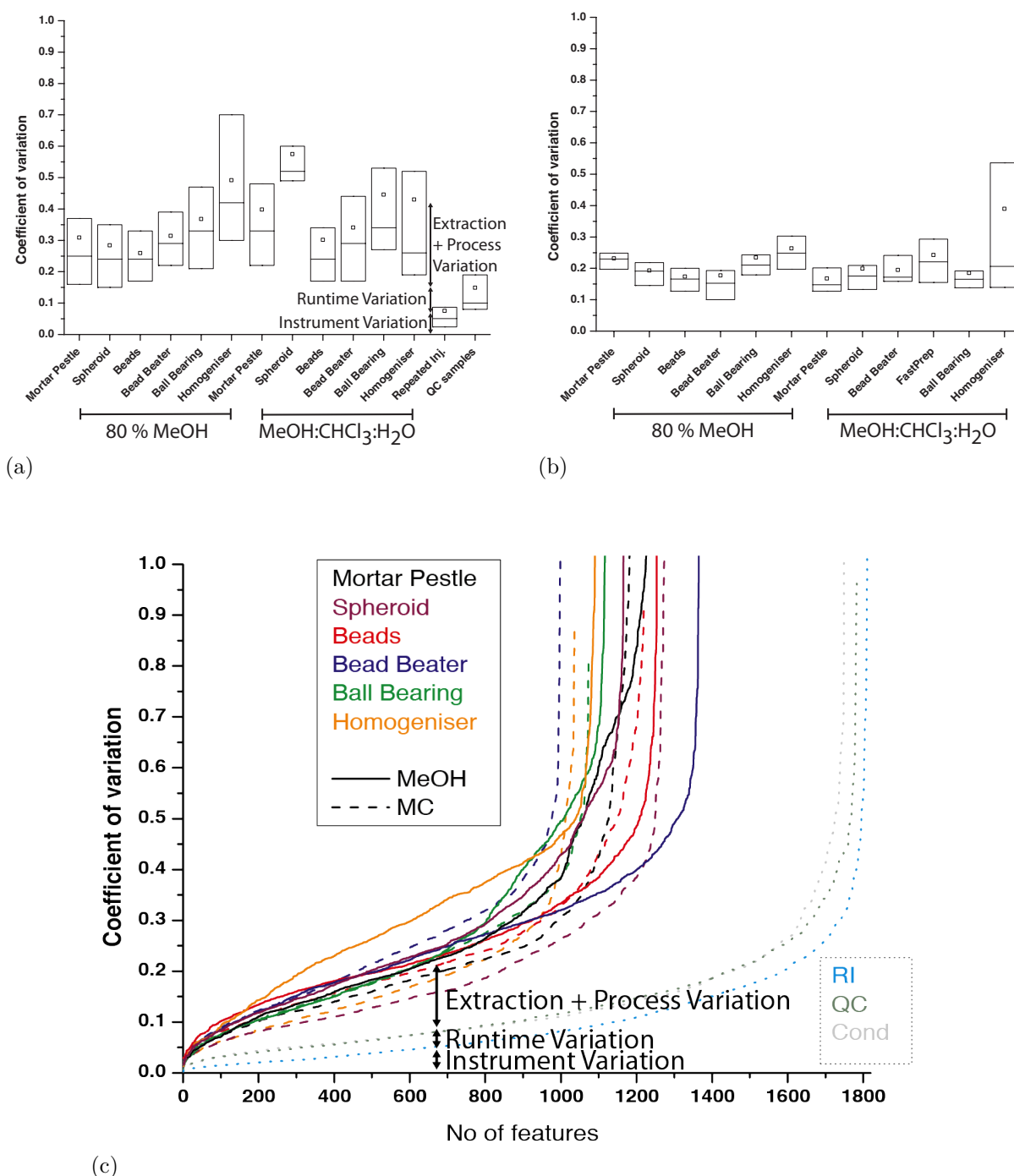


Figure II.6.: **Precision of extracted metabolite features:**

The coefficient of variation (CV) is used to assess the precision for the metabolite features of each extraction method for GC-MS (II.6a), NMR (II.6b) and LC-MS (II.6c). For GC-MS and NMR features the CV distribution is presented as box plots, whereas for a LC-MS the CV is ranked in ascending order. The box spans between the 25th and 75th percentile, with a line at median height and a dot representing the mean.

different subset of features, which were combined in the QC. Most likely, a combination of those two phenomena might be the case. The most interesting observation, however, is how the individual types of variation partition and correlate, especially for the LC-MS experiment. Repeated injection and QC sample CVs rise with the same slope, but with a small and constant offset. This offset, the difference between throughout run variation (QC) and machine variation (repeated injection), equals the pure runtime variation (2 - 3%). The fact this offset is constant suggests that, in a well-performed LC-MS experiment (*i.e.* no intensity drift due to source contamination), runtime variation is constant for each feature. The pure technical reproducibility, however, seems to depend on the feature, as its CV rises with the number of features. Thus, in this experiment, the CV of a feature only depends on its ionisation efficiency, which seems to be constant throughout the run.

Generally, nearly all methods yielded a median CV of 15% - 30% across the three platforms. This is in good agreement with values published for several GC-MS and LC-MS extraction benchmarks [615, 572, 394]. For NMR however, the median CV was higher than the ones reported in literature [349, 314, 35]. Usually though, the NMR-derived instrument variation is much less than the variation between two biological replicates [71], even taking metabolite fitting errors into account. Therefore the high amount of variation must reflect the substantial amount of process variation.

Some individual method/solvent/platform combinations (e.g. spheroid/MCW extraction for GC-MS; homogeniser/M80 for LC-MS; and homogeniser/MCW for NMR) had a CV distribution clearly skewed towards a poor performance. Similar unexplainable method bias for the same extraction protocol on different platform was previously described by t'Kind *et al.*[572]. The spheroid/MCW results are particularly intriguing. For NMR, the median CV lies well within the other methods; for GC-MS, it clearly performed worse than the other methods; whereas judging by the LC-MS data, this method had the lowest CVs throughout.

II.4.6. Yield

Nearly all extraction comparisons discuss coverage and reproducibility. However, yield, the total amount of extracted metabolites, is rarely considered, but should be equally as important. This is clearly the case for the current study, which has a tough matrix with

different cell types, and it is unclear which disruption method does the best job at reaching all compartments. The problem is the lack of clear metrics. Recovery, *i.e.* what percentage of a solubilised metabolite is recovered by the solvent, can be assessed by stable-isotope labelled metabolite spike-in [414]. The total amount inside the matrix, however, is unknown, which makes absolute yield difficult to assess. Therefore, a relative metric was adopted. For each metabolite feature, the most intense feature across all methods was set as maximum yield (100 %). The other methods intensities of this feature were scaled proportionally. The distribution of relative yields can then be plotted, as shown in figure II.7.

For GC-MS (figure II.7a), the median yield ranged from 50 % to 90 %. For LC-MS (figure II.7b) yield was between 50 % and 75 % and for NMR (figure II.7c) it ranged from 65 % to 95 %. Only LC-MS showed a solvent effect, with lower yields in the MCW samples. This goes in hand with the observation that M80 yielded 150 - 300 more peaks than their MCW counterparts. LC-MS had the highest within-method variance, but GC-MS and NMR had the highest between-method variance. Ball-bearing and homogeniser had the lowest overall yield. This might either be due to incomplete tissue disruption or, more likely, the difficulty of complete recovery from these devices, especially as those were the only methods that did not contain the sample in microcentrifuge/bead-beater tubes.

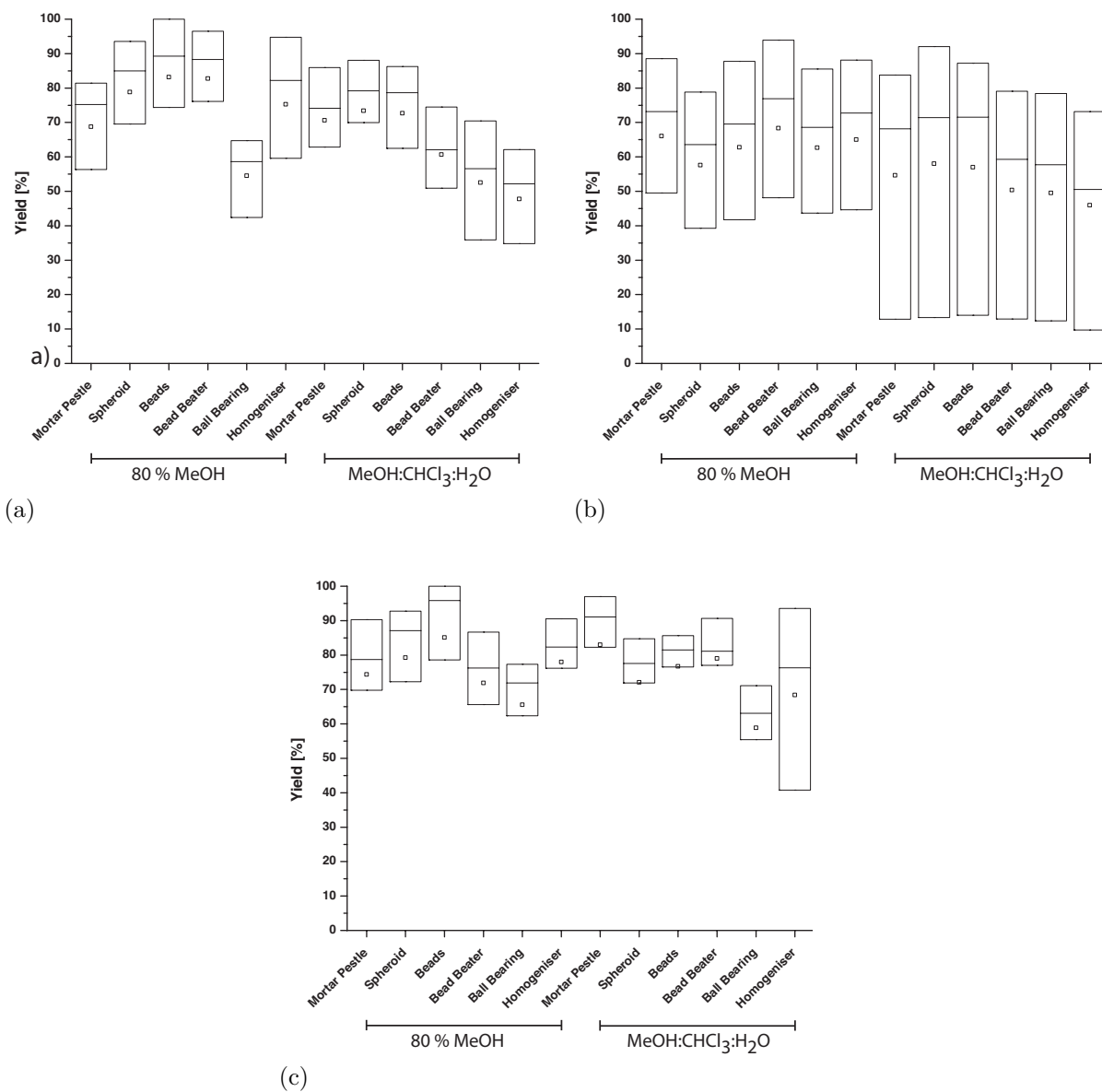


Figure II.7.: **Yield of extracted metabolite features:**

Yield for GC-MS (II.7a), LC-MS (II.7b) and NMR (II.7c) is assessed on a relative scale. The yield of any metabolite feature is expressed as a percentage of the highest signal intensity found by any of the extraction methods for this feature. The box represents the 25th and 75th percentiles, with a line at median height and a dot representing the mean.

II.5. Discussion

II.5.1. Necessity of a washing step during sampling

In the first step of the extraction optimisation, the *C. elegans* sampling protocol was put under scrutiny. The amount of bacterial contamination, stemming from *E. coli* as the worm's food source, was evaluated. This was done to assess whether an additional washing step of the worm pellet was required prior to extraction. Compared to the NMR signal of the extracted worm pellet, the same volume of "used" buffer, with suspended bacteria, showed no appreciable contribution of bacterial biomass. Close inspection revealed there were a few low-intensity peaks. However, these were negligible compared to the wealth of peaks in the worm spectrum. Furthermore it must be noted that only large (L4 or adult) worms had settled after 5 min. As this experiment was performed using a mixed population, there is a fair chance that some of the bacterial signal was partially due to the presence of smaller worms in the volume above the pellet. The negligible impact of bacteria is not surprising. Worms are usually sampled shortly before exhaustion of the bacterial lawn, which also firmly "sticks" to the plate throughout the harvesting process. Based on this result, future experiments, including the extraction comparison, were conducted without a pellet washing step.

II.5.2. What is the "best" method

Some methods were best on on a particular platform, even though under-performing when analysed using a different analytical platform. Thus, definition of the apparent "best" method depends in part on which analytical platform is used for evaluation. In fact, several disruption methods, such mortar and pestle, bead-beating, and grinding with beads in a reciprocating mill, performed consistently well. However, it was decided to select a single optimal extraction method for subsequent studies. Bead-beating performed consistently well with respect to the total number of metabolites (or metabolite features) detected, precision, and yield. It also can be used for high-throughput (≤ 24 samples per 3 min) applications. Most importantly, however, compared to grinding with a mortar and pestle, it is independent of the operator and therefore results should be reproducible across labs. Overall, tissue disruption methods had a lower impact on the result than solvent systems - at least for GC-MS and LC-MS. 80 % methanol rather than chloroform/methanol was

chosen as the best solvent system. It is simpler than the chloroform/methanol protocol that requires phase separation, metabolite partitioning, and accessing a single phase later., and so it is likely to be more robust for high-throughput studies in particular. Also, it is able to extract a variety of lipids, which can be analysed via LC-MS and reduce the need for a separate, lipidomic approach.

The final sampling and extraction protocol, using 80 % methanol and bead-beating, which will be used for all other experiments of this thesis, is outlined in the appendix (B2).

II.5.3. Observed metabolites - In comparison to previously published studies

When this study was initiated, only Blaise *et al.*[50] and Atherton *et al.*[23] had published *C. elegans* metabolite annotations for NMR and GC-MS, respectively. All metabolites assigned by this study using NMR and GC-MS have been confirmed in the meantime, by other publications [51, 299, 392, 86, 288, 76, 15, 75, 85]. However, this is hardly surprising, as those are all abundant and frequently observed metabolites from central carbon and energy metabolism, including amino acids, small organic acids, (sugar) nucleotides and sugars.

Nevertheless, [50, 23] are comparable with regards to sample generation and analytical platforms, also reported more uncommon metabolites that were not observed in this study. These include trimethylamine-N-oxide (TMAO) (both Atherton *et al.* and Blaise *et al.*); 4-hydroxyphenylacetic acid, α -ketoglutarate, β -alanine, phenylacetylglutamine, indoxyl sulphate, fructose, and malonate (Atherton *et al.*); and β -hydroxyvalerate, penicillamine, saccharopine, 2-keto-3-methylvalerate, N-acetylglutamate, isovalerate, homocysteine, homocarnosine, 5-methoxytryptamine, dihydroxyacetone, folate, galactose, indole, phenylacetate, β -phenylpyruvate, anserine, and picolinate (Blaise *et al.*). Several of those are common mammalian urinary metabolites, unlikely to be found in tissues (TMAO, phenylacetylglutamate, phenylacetylglutamine, phenylpyruvate). Some are not likely to be present in NMR-detectable concentrations (*e.g.* indole, folate, 5-methoxytryptamine, saccharopine, dihydroxyacetone, picolinate). Penicillamine, as a penicillin metabolite, is very unlikely to be present. Others are not implausible (β -hydroxyvalerate, 2-keto-3-methylvalerate, N-acetylglutamate, isovalerate, homocysteine, homocarnosine, galactose, anserine), but have

not been observed in this or subsequent studies [660] (others in preparation). It appears that many previously reported *C. elegans* metabolites are likely to be misassignments. There are still several resonances and peaks that remain unassigned in the NMR and GC spectra, respectively. This shows the need for improved metabolite assignment tools, as well as the potential strength of untargeted approaches.

II.5.4. What is limiting the coverage and assignment?

Metabolomics, as an untargeted profiling approach, heavily relies on two parameters. Firstly, "as wide as possible" metabolite coverage, and secondly, the ability to annotate biomarker candidates. To this end, this chapter benchmarked extraction methods and established standard profiling workflows for NMR, GC-MS and LC-MS. The coverage and ease of annotation was highly dependent on the analytical detection method. Therefore it is important to understand limitations of the three profiling platforms used.

NMR spectroscopy yielded the least number of metabolites. However, those were readily assigned. Assignment for common metabolites at least is straightforward, as informative 2-D experiments (COSY, HSQC, J-Res) can be run. HSQC spectra are particularly helpful as they connect protons (1 ppm to 9 ppm) to the much wider spectral range of carbon (10 ppm to >200 ppm). The reason why not more than 32 compounds could be annotated with confidence was because of the low inherent sensitivity of HSQC spectra. This is due to its four times lower gyromagnetic ratio, but more importantly due to the low natural ^{13}C abundance of only 1.1 %. Another problem, especially in the ^1H domain, is the significant overlap of spectral resonances, as no prior separation is used. Particularly in the region between 2.5 ppm and 4.5 ppm it is not always apparent whether the signal represents a complex spin system, a spin system overlap or a combination thereof. It is very likely that some low-intensity metabolites are not detected because of peak overlap problems, even though they would be above absolute detection limits in pure solutions.

GC-MS coverage and annotation complexity falls somewhere between NMR and LC-MS. A retention-time-locked spectral database yields quick annotation success. However, even the best fragment matching systems cannot outperform experienced spectrometrists [140, 25, 434]. Therefore it is vital to confirm database hits to reduce the number of false positives, *i.e.* falsely annotated peaks. Done carefully, this is a very robust workflow [38]. The workflow implementation used in this study, however, has a major drawback, never-

theless. It is not truly untargeted. For the used software package (GaVin) to integrate a metabolite, this metabolite must be present in the database, and therefore a high false negative rate, can be expected. More uncommon, or even unknown and therefore especially interesting, metabolites are likely to be missed. This is even more likely to be the case with *C. elegans*, as there is fewer metabolome resources available compared with *e.g.* humans. Indeed, under close inspection, the total ion chromatogram (TIC) displays an abundance of peaks currently not annotated and therefore not considered by the employed workflow.

LC-MS unarguably has the highest coverage, but annotation is difficult. Even though electrospray ionisation (ESI) mass spectrometry based metabolite databases have been growing substantially over the last few years [530, 630, 251], they are still very far from being comprehensive [320, 321]. However, before a database may be queried, it must be decided whether the candidate mass is the charged metabolite itself or an adduct or even an isotope. To some degree this can be done by rule and correlation based methods, such as the CAMERA software package [338]. Once the exact $M^{+/-}$ is known, but no database entry exists, there are plenty of possible molecular formulae, albeit the ever improving mass accuracy of modern high resolution mass spectrometers ($\approx 1\text{ ppm}$). Furthermore, there is also growing evidence that a vast majority of the peaks observed in LC-MS are not of biological origin, but rather contaminants from the sampling, processing, consumables or analytical platform [31, 194]. Altogether, these problems severely affect the advantages gained from the broad LC-MS coverage.

The following chapter (III.) tries to address the discussed limitations for NMR and GC-MS. Analytical and processing approaches more advanced than the standard workflow will be developed. Chapter IV. addresses the mentioned problems with LC-MS data. First, by *in-silico* experiments and algorithm development, but with a wet-lab based proof-of-concept following. The improved methods are then used for the final and major biological experiment - the longevity mutant screen in chapter V..

III.. Methods to increase the metabolite coverage and annotation on NMR and GC-MS based profiling platforms

Aims and Ojectives:

- Increase sensitivity for HSQC experiments and perform automatic annotation
- Increase NMR coverage through a chromatographic separation step
- Develop a workflow to analyse GC-MS data without *a priori* metabolite information

III.1. Introduction

In the last chapter, in section II.5.4, I discussed limitations for the three different profiling platforms. This chapter proposes solutions for some of the constraints with application to *C. elegans* samples. Section III.2 will investigate the feasibility of automatic database-assisted metabolite assignment of an isotope-enriched *C. elegans* sample, acquired using a 2D HSQC pulse sequence that has not previously been used for metabolomics. Section III.3 will explore the advantage of using fully deuterated ion pair chromatography, to separate a high concentration *C. elegans* extract for online (LC-NMR) and offline NMR detection. Section III.4 will analyse a dataset of long-lived and short-lived *C. elegans* mitochondrial respiratory chain mutations by a workflow based on TICs following retention time alignment.

The focus of this chapter is not to exhaustively test methods or perform in-depth biological characterisation, but rather to show that many of the constraints of the analytical platforms can be addressed by simple means.

All of the projects in this chapters were performed in a collaborative effort.

The isotope enriched sample of section III.2 was prepared and acquired together with Yevgeniy Izraelit from Cornell University (Ithaca, NY, USA). The high concentration sample of experiment III.3 was also prepared by Yevgeniy Izraelit at the lab of Frank Schroeder at Cornell University. Section III.4 is based on a collaboration with the lab of Aleksandra Trifunovic at the University of Cologne (Germany). Worm pellets were generated by Marija Sumakovic at the University of Cologne.

III.2. HSQC sensitivity improvement by universal ^{13}C labelling

III.2.1. Background and introduction

Since introduced by Bodenhausen and Ruben in 1980 [55], heteronuclear single quantum coherence (HSQC) NMR experiments have become indispensable for structural assignments and elucidation, and in particular ^{13}C - ^1H HSQC experiments [643]. However, its major drawback is its low inherent sensitivity and thus long acquisition time and poor signal to noise ratio. Therefore it mostly yields information on high abundance metabolites; which often are well known already. Interesting, but less intense peaks remain difficult to assign.

The low sensitivity is due to two factors. One is the four times lower gyromagnetic ratio and thus excited energy level of the ^{13}C nuclei, compared to ^1H at the same magnetic field strength; but the main factor is the low natural abundance of only 1.1 % ^{13}C . In order to follow a protein's peptide bond backbone, NMR-based structural biology routinely has to use three dimensional ^1H - ^{13}C - ^{15}N sequences. With a 0.37 % ^{15}N natural abundance, ^{13}C to ^{15}N magnetisation transfers become rare events. In order to avoid week long acquisition times, the structural biology field has developed several strategies.

A simple and often applied strategy to increase low natural abundance is artificial isotope enrichment [184]. Indeed, as many protein expression system, such as *E. coli* can be grown on minimal media this is readily possible; with ^{13}C glucose as sole carbon and ^{15}N ammonium salts as sole nitrogen source [218].

With full ^{13}C abundance, carbon - carbon couplings, which at natural abundance are rare events, become an issue. Classic broadband decoupling techniques used to suppress ^1H - ^{13}C satellites do not work for ^{13}C - ^{13}C coupling at full ^{13}C incorporation [605], and would lead to unacceptable sample heating. To this end, the structural NMR community has developed pulse sequences that circumvent C-C couplings. In particular, the constant time HSQC (ct-HSQC), proposed in 1992 by Vuister and Bax [605], is a pulse sequence that is often used for fully labelled multidimensional protein NMR spectroscopy [391]. A ct-HSQC pulse sequence is very similar to a standard HSQC sequence. However, instead of the usual variable ^{13}C evolution period t_1 , a constant time is used. Also during this interval, an ad-

ditional 180° pulse refocuses ^{13}C - ^{13}C coupling constants. Therefore couplings will not split but appear as cross-peaks along the F_1 domain and (ideally) not interfere with the rest of the spectrum.

So far only a few studies have used stable isotope labelling to improve HSQC sensitivity for small molecules. In 2004 Kikuchi *et al.*[316] were the first to ^{13}C and ^{15}N label *Arabidopsis thaliana* plants to probe ethanol stress response and nitrogen fluxes during germination, respectively; both with HSQC pulse sequences. In 2008 Chikayama *et al.*[94] built an online-database of > 1000 ^{13}C - ^1H HSQC chemical shifts they had assigned to 142 metabolites from *Arabidopsis thaliana* and silkworm larvae (*Bombyx mori*). In 2010 they went on to refine their approach by implementing an algorithm that automatically considers multiple HSQC resonances and assigns plausible metabolites. This assignment is based on a statistical likelihood value [93]. Gowda *et al.*[206] extended the use of stable isotopes in conjunction with HSQC sequences to matrices that cannot readily be isotope enriched. By derivatising blood plasma samples with ^{13}C formate and ^{15}N ethanolamine they were able to show a significant improvement in HSQC sensitivity. However, this approach only covers part of the metabolome, is only useful for identification when chemical standards are available and does not help with de-novo assignments. In 2012, when this study was already completed, An *et al.*[15] demonstrated the feasibility of HSQC profiling with ^{13}C enriched *C. elegans* samples. However, none of these studies used a constant time HSQC pulse sequence, even though this would be the structural biology community's gold standard approach for isotope enriched samples.

Therefore, this experiment aims to evaluate, whether ^{13}C labelling and the use of a ct-HSQC pulse sequence is a valuable approach to improve annotation of *C. elegans* HSQC visible metabolites.

III.2.2. Materials and Methods

III.2.2.1. Sample generation

An unlabelled control sample was prepared by growing wild-type N2 worms according to standard procedures, outlined in the appendix (B1.1). To generate the ^{13}C enriched *C. elegans* samples, a protocol developed in chapter IV. was used. Agar plates were prepared without bacterial nutrient (peptone). Instead, the pellet of a 25 ml overnight culture of the *E. coli* K-12 wild type strain NCM3722 was spread onto the agar plates. NCM3722 was grown in a minimal medium, with ^{13}C -glucose and ^{14}N -ammonium chloride as sole carbon and nitrogen source, respectively. To allow full isotope enrichment, worms were grown over two generations on the isotope medium. For detail on this method, please refer to section IV.2.1.4.

All worm cultures (isotope enriched and unlabeled) were age synchronised by the method described in B1.2. Worms were harvested as young adults and extracted by bead-beating in 80% methanol, according to a consensus protocol based on the extraction comparison in chapter II.. For experimental detail, please refer to B2.

In both cases, the biomass of three plates of worms were pooled to yield one biological replicate, in order to improve the signal to noise ratio.

III.2.2.2. Acquisition and processing

For NMR spectroscopy, the samples were prepared according to the general procedure described in B3.1. Acquisition was performed at 300° K on Bruker Avance II spectrometer at 800 MHz ^1H observation frequency. A cryogenically cooled triple-axis inverse probe was used. The system was controlled by Bruker TopSpin 2.1. For the regular HSQC spectra a Bruker gradient HSQC sequence with pre-saturation and sensitivity enhancement was used. The constant time sequence (ct-HSQC) used the Bruker "hsqcctetgpcsp" sequence, adapted from Isaacson *et al.*[273] and originally based on Vuister and Bax [605]. The constant t_1 evolution time needs to be optimised, to target the coupling constant that needs to be "removed" during the 180° refocusing pulse. As most metabolites of a complex mixture have at least some resonances in the aliphatic region, it was decided to target those. Aliphatic C-C couplings are usually around 32-40 Hz. Therefore the constant time (d_{23})

was set to ($2T = 1/J = 1/38.5 \text{ Hz} = 26.6 \text{ ms}$). The pulses in the carbon domain were adiabatic, shaped pulses [564]. For both the HSQC and ct-HSQC the 90° pulse p_1 was set to 14 ms. After 32 dummy scans, 64 time domain transients with 2048 increments in the F_2 domain and 512 increments in the F_1 domain were acquired. The spectral sweep width spanned 10 ppm in the proton and 90 ppm in the carbon domain, centred at 4.70 and 90 ppm, respectively.

For processing inside TopSpin 3.1, the FID was multiplied by a Gaussian window function in the F_2 and a shifted sine-bell squared function in the F_1 dimension. The time domain spectrum was then Fourier transformed into a 1k by 1k data point frequency spectrum. Phase correction was performed by optimising the phase angle of 1D projections of selected peaks across the whole spectrum.

III.2.2.3. Spectral deconvolution and automatic database matching

For automatic assignment, the processed spectra were loaded into rNMR [361]. The signal-to-noise threshold was manually adjusted so that peaks had defined contours and no random noise was visible in spectral regions that are unlikely to contain peaks, *i.e.* the high proton, low carbon range. Peaks were detected using the default algorithm, without prior signal smoothing. In the constant time HSQC spectra, peaks stemming from the F_1 artefact were manually removed. For assignment, the HSQC reference spectra from HMDB were used. To this end all available spectra ($n = 669$) were downloaded (on the 21.11.2012) as XML files and read into R. Spectra were filtered by meta-information and only spectra recorded in similar experimental condition were kept (aqueous buffer with $pH \approx 7$)

As the reference spectra were recorded in a similar but not identical setup, ppm values can slightly deviate. This deviation might be composed of a systemic and an individual peak-by-peak component. In this experiment it was attempted to minimise both. The systemic shift was estimated by comparing ppm shifts of 15 peaks from different metabolites across the spectrum, that were assigned in chapter II., with the ppm values of the corresponding reference peaks.

Peak to peak deviations are more difficult to deal with. When matching the spectra against a reference, the acceptable error must be large enough to allow for the peak to peak de-

viation. However, when the allowed error margin is too large, one peak may match other neighbouring peaks, making the match ambiguous and therefore reducing the power of the query. Therefore, it is important to adapt the error tolerance for each dataset, when an automatic search algorithm is used [93]. The best way to find this value would be to annotate most metabolites in the mixture and determine their (median) distance from the reference. However, this tedious approach is contradictory to the intention behind automatic assignment, which the goal of this experiment. A Monte-Carlo simulation was used to optimize these parameters for both proton and carbon shifts.

The selected cross peaks and the HMDB HSQC reference peaks were matched three thousand times. Each time, the allowed ppm error was randomly chosen within a window of 1 to 0.001 ppm and 2 to 0.005 ppm in the proton and carbon domain, respectively. After each iteration, it was counted (a) which fraction of metabolites from the HMDB reference spectra did not receive at least one match from a HSQC spectral peak. (b) how many HMDB metabolites were completely matched (each cross peak of the reference matched a cross peak of the spectrum). The convolution of (a) and (b) ($a * b$) was then calculated to determine the best trade off between the minimum (a) and maximum (b) matching window size, necessary to minimise false negatives and false positives, respectively. The maximum (best tradeoff) of the spectrum that resulted from the convolution determined the error margin for the final matching of HSQC peaks and HMDB reference spectra.

III.2.3. Results and Discussion

III.2.3.1. ^{13}C enrichment can achieve a drastic increase in sensitivity

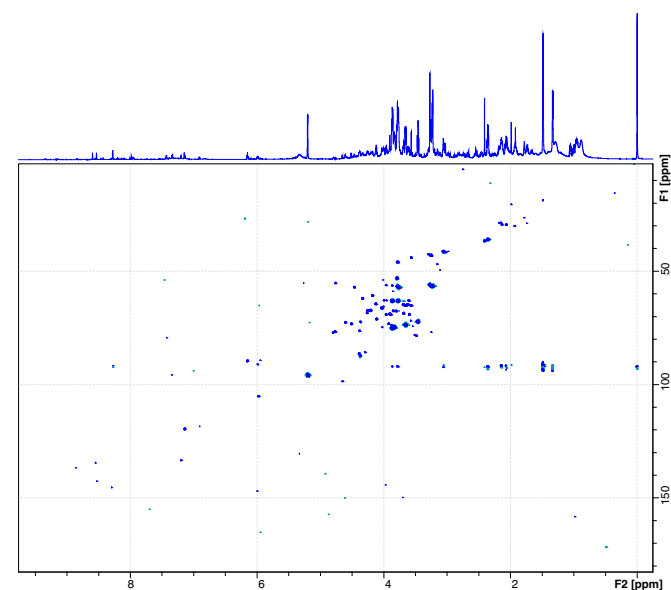
The first step in evaluating whether ^{13}C enriched samples acquired by ct-HSQC pulse sequences for *C. elegans* metabolite assignment are valuable, is to compare a fully ^{13}C labelled versus an unlabelled sample. To this end, wild type worms were either fed with *E. coli* grown on U- ^{13}C -glucose or unlabelled glucose minimal medium. D_2O reconstituted 80 % MeOH extracts of three pooled plates were then compared using ct-HSQC sequences. To make the spectra comparable, the signal-to-noise cut-off of both spectra was scaled such that the ^1H resonances at 5.20 ppm (from the anomeric protons of trehalose) matched each other's intensity. Results are shown in figure III.1.

As expected and documented in literature [316, 94], an increase in signal intensity obtained from isotope enrichment of the matrix is immediately obvious. Not only do cross peaks that are already visible in the ^{12}C extract (figure III.1a) become more intense, a plethora of new resonances appear above the detection limit (figure III.1b), especially in the otherwise sparse aromatic region; the aliphatic region also increases in resonance density. Resonances around $F_2 = 4$ ppm become crowded and overlap even in two dimensions. However, at a higher zoom level, displayed in figure III.2b, individual cross-peaks can still be well resolved. Increasing the number of increments in the F_2 domain would also be a readily implemented solution to reduce signal overlap.

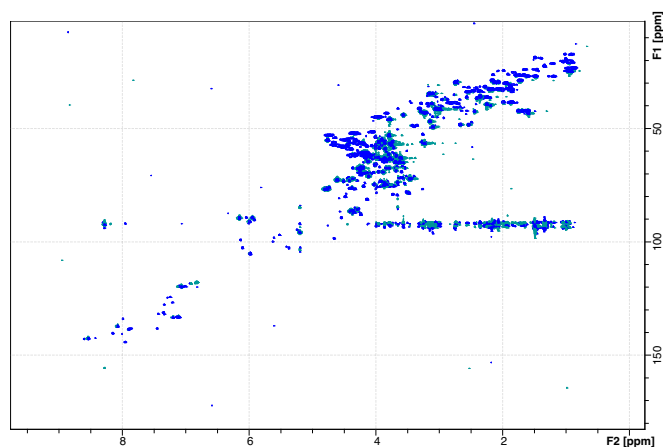
III.2.3.2. The constant time version outperforms the regular HSQC pulse sequence for ^{13}C enriched sample types.

As mentioned in the introduction, all HSQC spectra of ^{13}C enriched metabolite mixtures that have been reported in the literature were acquired with a standard HSQC pulse sequence, even though constant time HSQC sequences are more commonly used in structural biology. Therefore, both sequences were compared for their performance with small molecules (metabolomics).

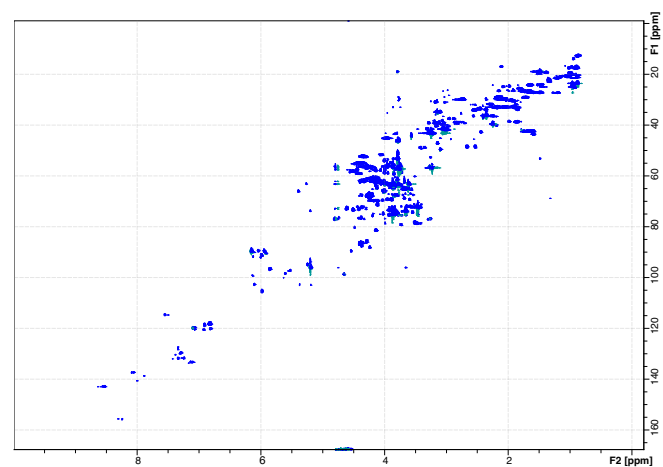
If the full spectra from the constant time (figure III.1b) and regular HSQC (figure III.1c) sequences are compared, there are not as many striking differences as between the acquisition of ^{12}C and ^{13}C samples. However, the ct-HSQC artefact parallel to the F_2 domain



(a)

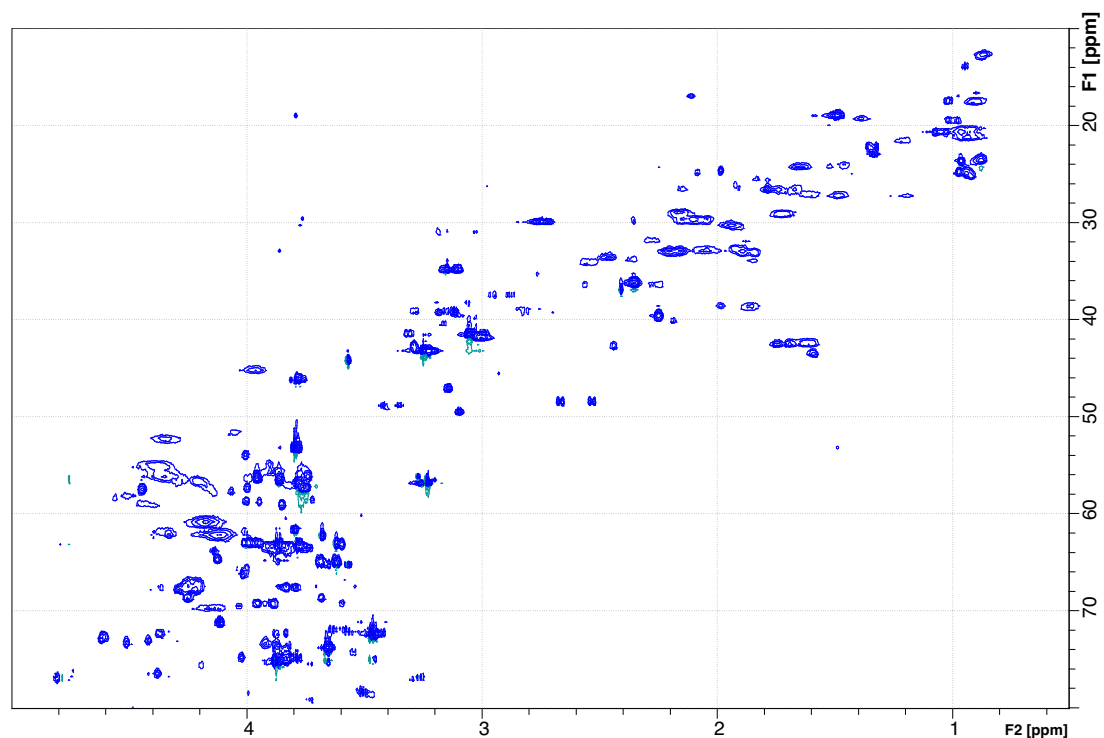


(b)

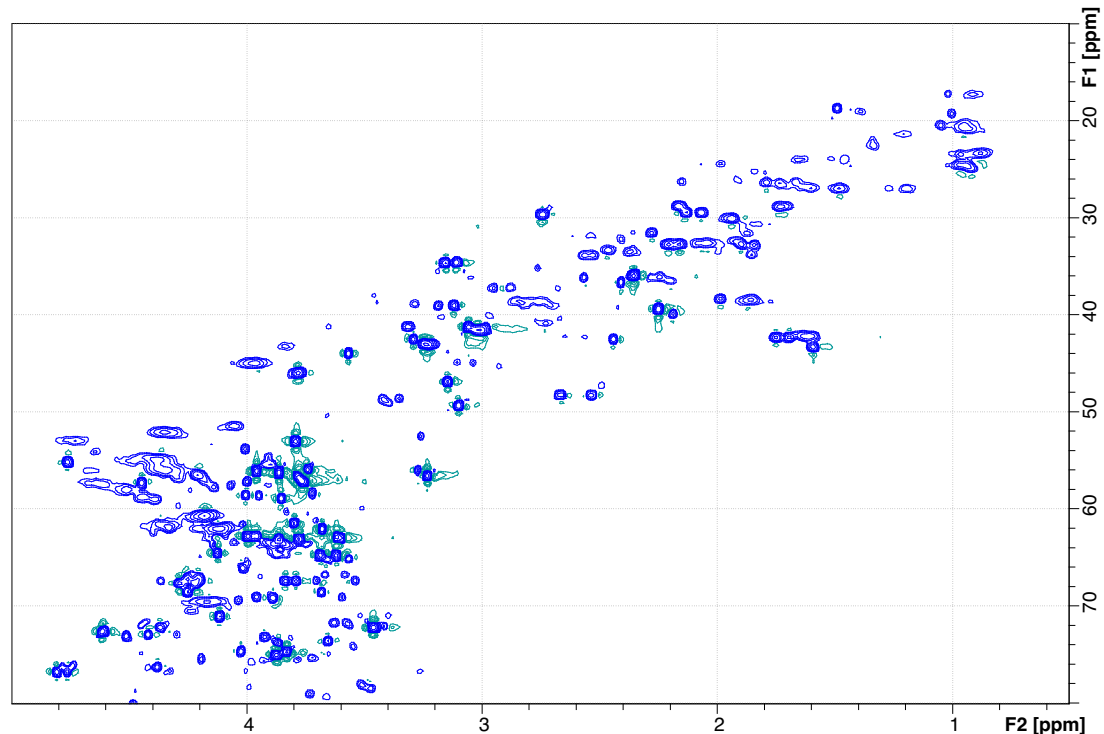


(c)

Figure III.1.: **Effects of ^{13}C enrichment on (ct)-HSQC spectra:** ct-HSQC spectrum of a natural abundance (III.1a) and ^{13}C enriched (III.1b) extract. Figure III.1c shows a regular HSQC sequence of the ^{13}C sample.



(a)



(b)

Figure III.2.: Difference between HSQC and ct-HSQC:

Magnification of the aliphatic resonances

($10 < ppm(^{13}\text{C}) < 80 - 0.5 < ppm(^1\text{H}) < 5$) of the ^{13}C sample, acquired with the regular HSQC sequence (III.2a) and the constant time HSQC sequence (III.2b).

at half F_1 shift range (90 ppm) is immediately obvious. Those are the C-C couplings, projected into this plane by the 180° refocusing pulse at the end of the constant time period. Luckily this artefact affects a region of the spectrum containing little or no information. At a higher magnification of the aliphatic region, the benefits of the constant time sequence become apparent. Compared to the cross peaks of the regular sequence (figure III.2a), the ct-HSQC's (figure III.2b) have an improved peak shape. There are many examples where carbon splitting patterns collapse into a single cross peak. Figure III.3 illustrates this phenomenon with two examples.

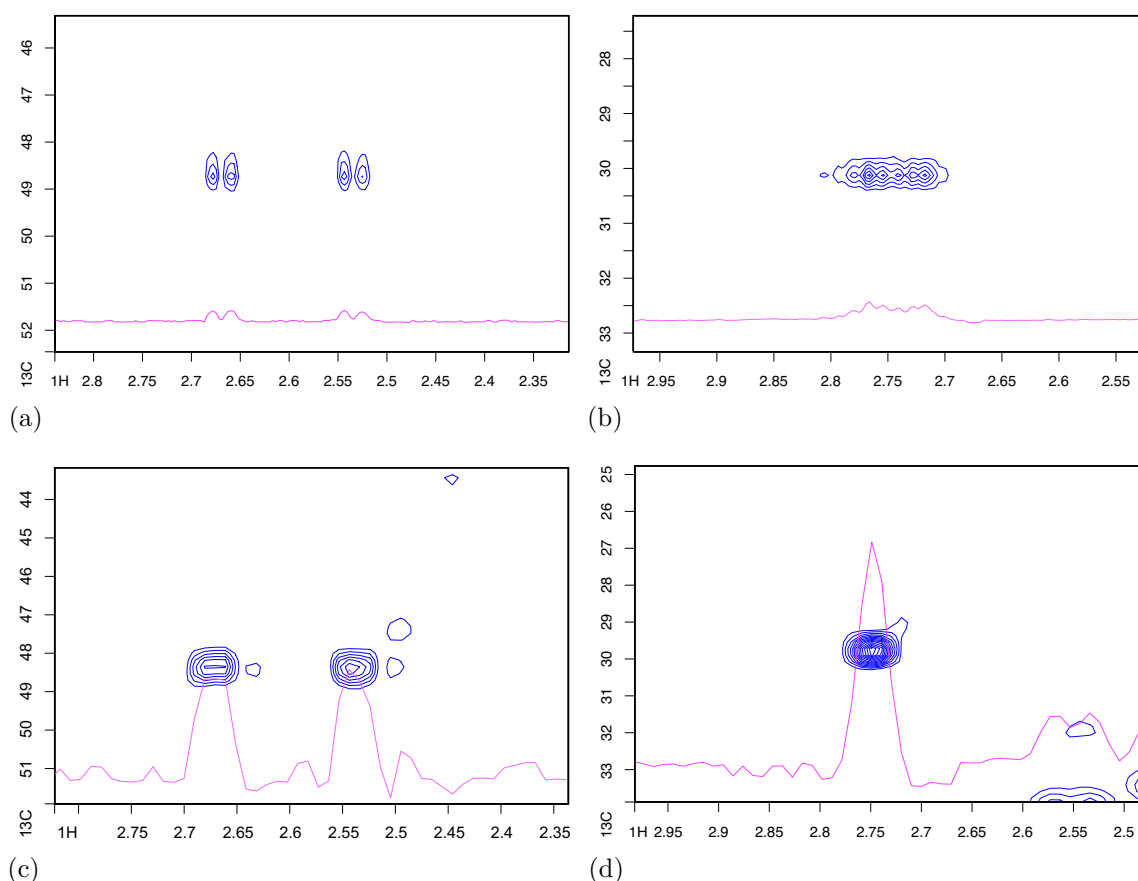


Figure III.3.: **Effect of the constant time HSQC sequence on C-C couplings:** Figure III.3a and III.3b show two cross peaks acquired by a regular HSQC pulse sequence. Figure III.3c and III.3d display the same cross peaks, but acquired with a constant time HSQC sequence. The magenta line below is the 1D projection of the 2D spectrum.

By removing C-C coupling, two doublets (figure III.3a) and one high-order multiplet (figure

III.3b) are collapsed into two (figure III.3c) and one (figure III.3d) singlets, respectively. Even though C-C coupling patterns may be of diagnostic use in the same way as H-H couplings are for one dimensional ^1H NMR spectroscopy, they might also lead to misinterpretation. It is not immediately apparent that the two example peaks acquired by the regular HSQC pulse sequence belong to the same carbon, rather than to several carbons inside similar chemical environments. Furthermore, the increased sensitivity, arising from the avoidance of C-C splitting, gives higher confidence for low intensity peaks - the ones HSQC spectroscopy usually struggles with and for which this project was initiated. Without a doubt, these properties of the ct-HSQC pulse sequence will ease HSQC based metabolite assignment, and may also improve automated assignment.

III.2.3.3. ^{13}C enriched ct-HSQC spectra are ideal for automatic compound annotation

Usually ^1H NMR metabolomics annotations are performed manually, using published assignment lists [442, 160] or more recently online databases, such as the human metabolome database [630]. Fully automated assignment would be highly desirable, considering the large number of metabolites present in complex mixtures and the resulting effort for an exhaustive compound annotation. However, automatic ^1H spectrum assignments are ambiguous, due to the small shift range and the use of only one dimension. Multi-dimensional spectra, such as HSQC spectra should allow for automatic annotation. Indeed, Chikayama *et al.* [94, 93] have built a database (SpinAssign) for automatic assignment of (^{13}C enriched) HSQC spectra.

The observed advantages of the constant time HSQC over the regular sequence, the simplification of the spectra by collapse of splitting patterns and the increased intensity of the resulting cross peaks, should improve automatic database-assisted compound annotation readily. As Chikayama *et al.*'s SpinAssign only contains ca. 250 metabolites and the quality of the reference spectra cannot be checked, it was decided to build a workflow using HMDB's HSQC reference spectra ($n = 669$). However, before any database query, the spectra had to be subject to a peak detection algorithm. Figure III.4 shows excerpts of the three spectra (^{12}C HSQC, ^{13}C HSQC, ^{13}C ct-HSQC) after peak detection with the R based software package rNMR [361].

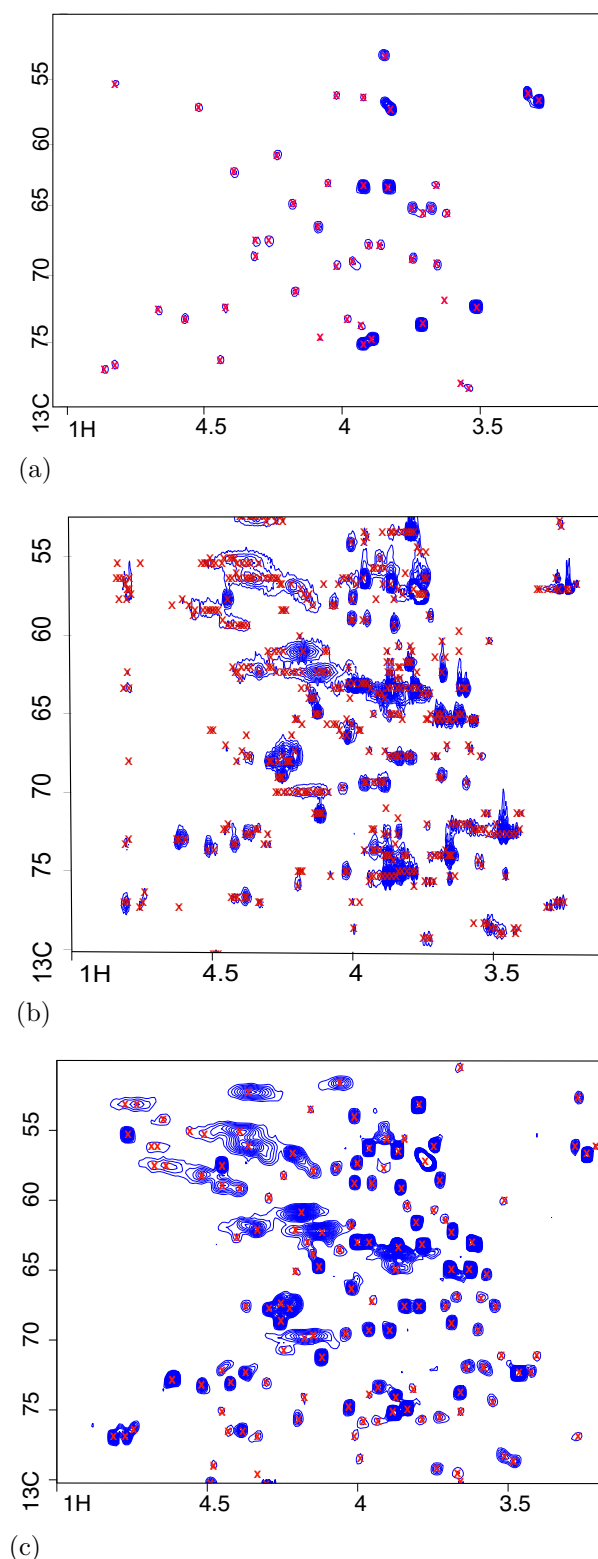


Figure III.4.: **Peaks detected in the ^{12}C and ^{13}C (constant time) HSQC spectra:** Spectral excerpts after peak detection with rNMR. Red crosses symbolise detected peaks in the ^{12}C (figure III.4a), ^{13}C HSQC (figure III.4b) and ^{13}C ct-HSQC (figure III.4c) spectra.

As expected, the ^{12}C HSQC showed the least number ($n = 69$) of detected peaks (figure III.4a). In the regular HSQC acquisition (figure III.4b) many more peaks ($n = 1242$) were detected than in the constant time version ($n = 340$) (figure III.4c). Checking which cross peaks were detected as apparent peaks showed a surprisingly good performance of the algorithm for the ct-HSQC spectrum. Each cross peak was detected once and only once. Even coinciding cross peaks was recognised by the algorithm, which then flagged all apices of a clustered peak group correctly. Conversely, for the regular ^{13}C HSQC acquisition, multiple apices were detected for a single cross peak, rendering the peak list completely useless. Close inspection revealed that the algorithm identified the local maxima of the C-C splitting patterns as individual peaks. Tweaking parameters, e.g. using smoothing filters, did not lead to noticeable improvement. Even though other software packages may perform slightly better, a ct-HSQC-like performance is unlikely, as differentiating between overlapping cross peaks and the maxima of complex splitting patterns is extremely challenging. Therefore, these results confirm the hypothesis that the constant time HSQC experiment is better suited for automatic assignment. For this reason it was decided to only compare the automatic annotation of the unlabelled and ^{13}C sample, acquired with the constant time HSQC pulse sequence.

In order to achieve the best database matching results, the HSQC spectra need to be aligned with the reference spectra as much as possible, by removing both systematic and random offsets. A constant chemical shift offset of 0.02 ppm in the proton and 0.228 ppm in the carbon domain could be removed. On average, this lowered the root mean square matching error between reference and acquired resonances by 3 fold in the proton domain and 0.5 fold in the carbon domain. A Monte-Carlo simulation was then used to determine the best error window for the HSQC - reference spectra matching. Results of this simulation are displayed in figure III.5.

The simulation results display the empirical relationship between the allowed error in the C and H domain and (a) the fraction of database metabolites yet completely unmatched by any HSQC cross peak (figure III.5a) and (b) the fraction of fully matched database metabolites, *i.e.* each cross peak of a reference metabolite is matched to a cross peak from the HSQC spectrum (figure III.5b). Interestingly, both functions' slopes were aligned diagonally, which means that to achieve the same number of matches, the error tolerance in the proton domain can be increased, if the error tolerance in the carbon domain is increased

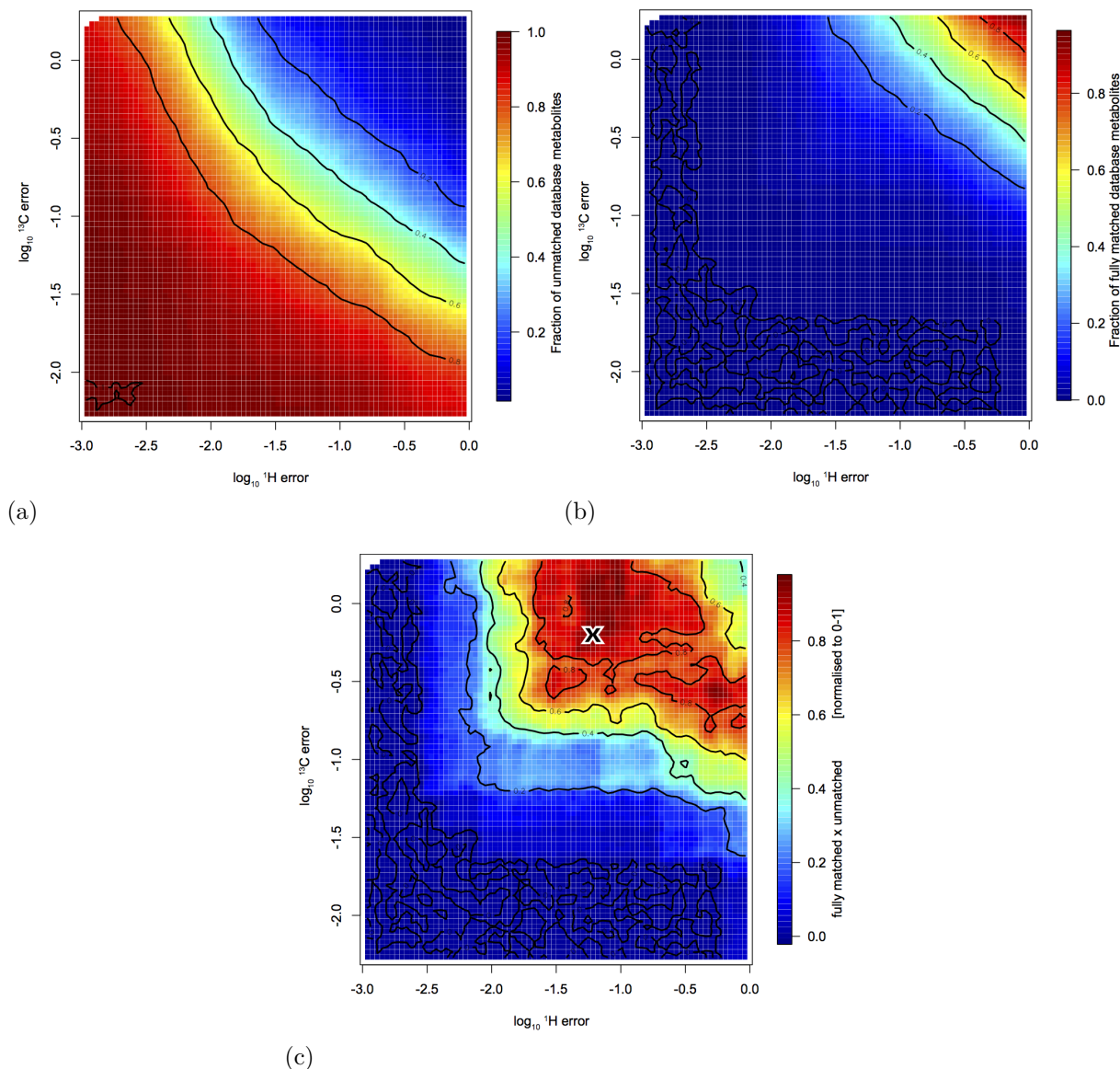


Figure III.5.: **Monte Carlo simulation to find the ideal matching error window:** HSQC peak - reference spectrum peak matching quality indicators in response to varied proton and carbon matching error. Figure III.5a shows the fraction of reference metabolites that were not yet matched with at least one HSQC cross peak. In figure III.5b the fraction of reference metabolites that were fully matched, *i.e.* each cross peak of the metabolite is matched to a cross peak in the HSQC spectra, is shown. Figure III.5c is the convolution of spectrum III.5a and III.5b. The "X" denotes the maximum of the convolution and therefore the best matching error window.

(and *vice versa*). In order to determine the best tradeoff between the danger of missing out on metabolites (a) and the risk of "over-assignment" (b), the optimum between those two functions was determined by convolution ($a * b = c$). The resulting spectrum (c), displayed in figure III.5c shows the recommended error ranges. It was decided to choose $\log_{10} = -1.25 \hat{=} \pm 0.056 \text{ ppm}$ and $\log_{10} = -0.25 \hat{=} \pm 0.56 \text{ ppm}$ as matching tolerance in the proton and carbon domain, respectively.

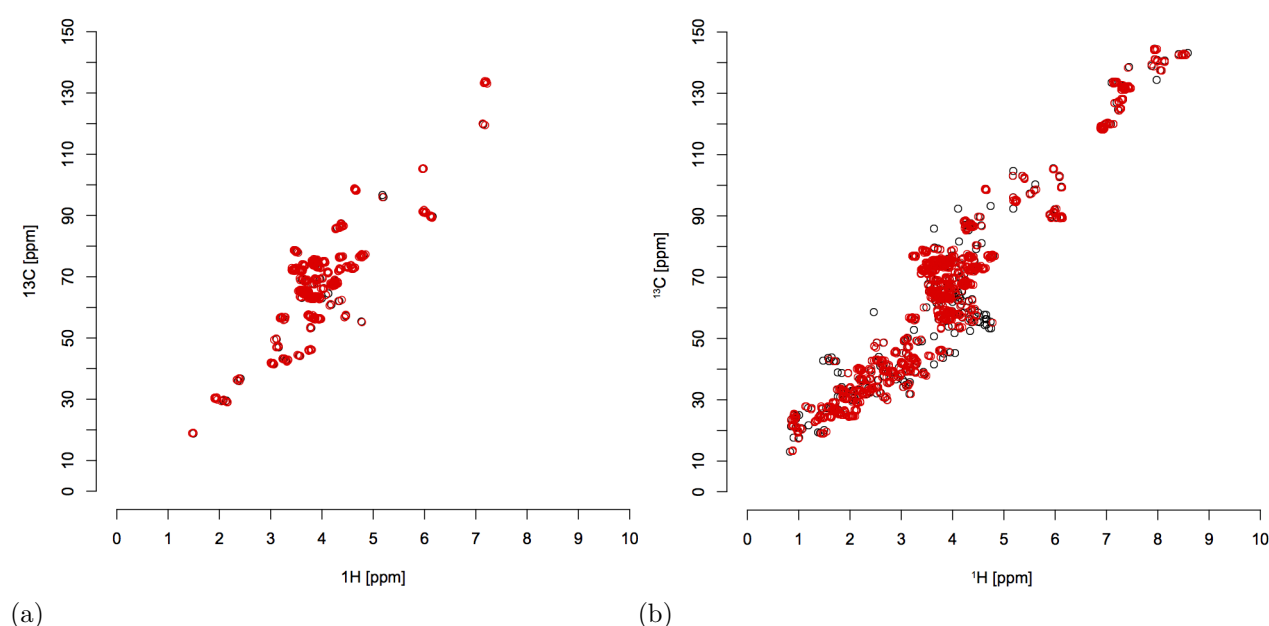


Figure III.6.: **HSQC cross peaks matched to HMDB metabolites:**

Peaks detected in the ^{12}C (III.6a) and ^{13}C (III.6b) ct-HSQC spectrum, that could be matched (red) or left unexplained (black) by cross peaks of the HMDB metabolite reference spectra using the prior determined matching error threshold.

Figure III.6 shows which peaks of the constant time HSQC spectrum of the ^{12}C (figure III.6a) and ^{13}C (figure III.6b) sample could be explained by HMDB metabolite cross peaks, using the determined error threshold. Even though most HSQC cross peaks were assigned to metabolites, a few still remained unassigned, particularly in the ^{13}C sample. This could be due to a too small error window, or more likely metabolites that are not recorded in the HMDB. *C. elegans* is known to produce a plethora of exotic NMR visible secondary metabolites, such as ascarosides [74, 536, 484].

The list of metabolites that were automatically assigned, in both the ^{12}C and ^{13}C ct-HSQC spectra, can be found in the appendix, in table C2.1. As expected, either more metabolites were detected in the ^{13}C sample, or detected with a higher confidence (more individual peaks of a metabolite matched). The number of metabolites for which all (100 %) peaks were matched to a peak in the reference HSQC spectrum increased by more than 7 fold with isotopic enrichment. If examined for their biochemical plausibility, almost all top (100 %) metabolite candidates appear to be reasonable suggestions for a cell extract. Suggestions with a lower confidence become more exotic, but completely implausible hits, for example drugs like metoprolol (No. 261) or theophylline (287), rank at the very bottom of the list. Furthermore, most of the metabolites manually assigned in chapter II. are recovered at the top of the list, indicating the robustness of this approach. However, there are also minor problems to be pointed out: the technique cannot differentiate between stereo descriptors. If enantiomeric pairs are listed in the database, those will be duplicated in the final list, as is the case for L/D - Glutamine (27/31) ⌀ although this could also be seen as more of a failure of database curation. Another potential pitfall is demonstrated by the assignment of citicoline (2). As an intermediate in the phosphatidylcholine biosynthesis pathway, it will generally be found in tissue extracts, but often below NMR detection limits. As it is made up of CDP and a choline subgroup, it is possible that the algorithm matched cross peaks to both individual compounds. To this end, all results will need to be manually validated.

However, as a first effort at assigning the metabolome of a new matrix, isotope enrichment, acquisition by a ct-HSQC sequence and database matching with optimised parameters, appears to be a valuable workflow. In chapter II., it would have saved a considerable effort and maybe led to the assignment of less commonly assigned metabolites - which was the major motivation for this chapter.

III.3. Online and offline LC-NMR separation of *C. elegans* extracts

III.3.1. Background and introduction

As discussed in chapter II., signal overlap poses a severe problem during the NMR analysis of complex mixtures, such as biofluids or tissue extracts. This is because NMR, in contrast to mass spectrometry, usually has no prior chromatographic analyte separation step. Particularly in a metabolite profiling experiment, where several samples need to be compared (semi-) quantitatively, LC with online-coupled NMR detection is of little use. Firstly, as NMR is an insensitive technique, it requires multiple experiments (scans) to be averaged to yield an acceptable S/N ratio. Secondly, NMR's strength in metabolomics is its fully quantitative nature and reproducibility. These would be compromised for on-flow LC-NMR. Higher costs, by an elevated demand for deuterated solvents and increased analysis times, are other prohibitive factors.

Nonetheless, LC-NMR is routinely used with complex mixtures for the purpose of compound characterization, most frequently in drug metabolism and natural product chemistry. However, the ability to exhaustively characterise a complex matrix by "pulling it apart" in a separation stage, is also useful for metabolomics. Biomarker candidates resulting from a metabolomics experiment or metabolites that are to be fitted for quantitative analyses, require assignment. Signal overlap in complex mixtures impedes assignment in two ways. Firstly, it is not always clear which resonances stem from the same compound and secondly, some similar compounds are not fully differentiable by proton NMR, *e.g.* ADP and ATP. The first problem may partially be addressed using covariance based bioinformatic approaches, such as STOCSY [106, 510] or CLASSY [498]. However, those techniques require many samples with varying concentrations and little overlap of the signals of interest. In order to address the second problem and to distinguish similar, but not ^1H NMR visible moieties, NMR spectroscopy of other NMR active nuclei, such as ^{13}C , ^{15}N and ^{31}P could be performed. However, this is not straightforward. It often requires isotope enrichment, appropriate NMR probes, expert knowledge and sometimes even special sample preparation [379, 380]. Therefore LC-NMR remains a viable option for the characterisation of new matrices and compounds.

LC-NMR is classically based on reversed phase chromatography. However, most NMR-visible metabolites of polar matrices, such as urine or cell extracts do not retain well on reversed phase [133, 525, 201]. To this end, other, more "polar friendly" types of chromatography have been evaluated for LC-NMR. Those range from hydrophobic interaction (HILIC) [201], to ion exchange [133, 207] up to ion pairing [525] chromatography. Even though all of them show superior performance for highly polar metabolites, ion pair chromatography can be advantageous as it does not require specialised chromatography columns like HILIC or ion exchange [87]. It only requires a reversed phase system, to which an ion pair reagent is added. Ion pairs consist of a positive or negative charge that are linked to a hydrophobic moiety. The charge will interact with oppositely charged compounds and the hydrophobic moiety will retain on the phase, until eluting with organic solvents. Most contemporary ion pairing methods have been developed with tertiary and quaternary butyl amines, as those retain a plethora of central carbon metabolism metabolites [525, 378, 375, 68].

The method for LC-NMR developed by Simpson *et al.*[525] uses a tetrabutylammonium salt. They report an effective separation of the polar constituents of complex environmental samples. However, they also noted that the ion pair's butyl resonances impedes interpretation of several aliphatic regions and requires heavy signal suppression of the affected ppm regions.

For this reason, this experiment's purpose is to evaluate the utility of a fully deuterated ion pair, such as d_{27} -tributylamine, that only requires standard solvent suppression. It could be a valuable complement to online and offline NMR detected reversed phase separation for the characterisation of small, polar molecules in a *C. elegans* tissue extract.

III.3.2. Materials and Methods

III.3.2.1. Sample generation

The *C. elegans* wild type sample was prepared by Yevgeniy Izraelit at Frank Schroeder's lab at Cornell University, Ithaca, NY.

To obtain a highly concentrated extract, worms were grown in liquid culture, rather than agar plates, as this method is the most appropriate for the generation of high amounts of biomass. Culturing was performed according to a standard procedure described in the wormbook [544]. The culture of approximately 1 million worms (100 plate equivalents) was frozen and freeze dried before posting the sample to Imperial College London.

The sample was extracted using the 80 % methanol bead-beater protocol, described in the appendix's protocol section (B2). Two consecutive extractions of 5 ml solvent were performed. Three aliquots of equal volume were made up, dried and stored at -80° until analysis.

III.3.2.2. Instrumentation and acquisition

For liquid chromatography an Agilent 1100 HPLC system (Waldbronn, Germany) system with a quaternary pump, Agilent sample manager, Agilent VWD UV detector set to measure at 254 nm, and a vacuum degasser was used. Separation was performed on a ACE 5 C18-HL column. This 150 mm by 4.6 mm C18 column with 5 μ m particle size has a higher loading capacity than conventional C18 columns. Thus, it allows injection of concentrated samples. Hystar software (Bruker Daltonik, Bremen, Germany) was used to control the whole LC system. The gradient for reversed and ion pairing chromatography can be seen in table III.1 and III.2, respectively.

Whereas the reversed phase method is a linear gradient, the ion pairing method mainly uses isocratic elution profiles at different organic content. The later is based on Luo *et al.*[378], as they optimised their method on the same LC system as the one used for this study. For both methods, solvent A was deuterium oxide (Sigma) and solvent B acetonitrile, both with 1 % acetic acid added as modifier. In the ion pairing chromatography 10 mM tributylamine was added and dissolved by sonication for 15 minutes. For the deuterated online and offline chromatography, d_3 - acetonitrile (Fisher), d_4 - acetic acid (Sigma) and d_{27} -

Time	A[%]	B[%]
0.00	97.0	3.0
5.00	97.0	3.0
40.00	0	100
50.00	0	100

Table III.1.: Reversed phase gradient

Time	A[%]	B[%]
0.00	97.0	3.0
15.00	97.0	3.0
25.00	80.0	20.0
55.00	80.0	20.0
60.00	65.0	35.0
65.00	65.0	35.0
70.00	40.0	60.0
75.00	40.0	60.0
75.10	10.0	90.0
85.00	10.0	90.0

Table III.2.: Ion pairing gradient

tributylamine (Qmx isotopes, Thaxted, UK) were used. The system was equilibrated for at least 15 minutes, before injecting 50 μ l into a solvent flow rate of 0.2 ml per minute.

For the online separation the LC eluent was directly coupled to a 60 μ l Bruker flow cell probe. This flow cell probe was inside a Bruker Avance III DRX600 spectrometer, with a 14.1 T magnet, operating at 600.13 Mhz ^1H observation frequency and controlled by TopSpin 3.1. Bruker's "lc2grdonflow" pulse sequence was used and set up with the provided GUI. This sequence is based on a standard NOESY sequence, with an intermediate preparation scan prior to the acquisition scans to adjust receiver gain and solvent suppression. Solvent suppression was performed by a shaped pulse WET sequence [451, 529], which was followed by ^{13}C broadband decoupling prior to the acquisition. During the acquisition, eight time domain transients were collected over a sweep with of 14 ppm and with a FID resolution of 8k and then added, to yield one spectrum every ≈ 10 seconds. The 90° pulse width was determined by manually introducing an aliquot of the sample, diluted in a solvent mixture reflecting starting conditions, into the flow probe. During processing in TopSpin 3.1, the individual time-domain spectra were multiplied with a 1 Hz sine bell window function with a sine bell shift (SSB) of 2.5, prior to Fourier transform. After phase correction at multiple retention times, the ^1H domain was calibrated to the anomeric proton of trehalose at 5.20 ppm.

The off-flow acquisition was performed in two steps. First, manual collection of one minute LC eluent fractions with subsequent solvent evaporation. In the second step, the fractions were prepared (see B3.1) and analysed by conventional tube NMR [37]. Spectra were

acquired on a Bruker Avance II spectrometer with a 14.1 T magnet (Bruker Biospin; Rheinstetten, Germany) operating at 600.13 MHz ^1H observation frequency. Due to the high number of samples a BACS automatic sample changer (Bruker) was used. After introduction into the magnet, sample temperature was allowed to settle to the 300°K of the probe. Automatic shimming and receiver gain adjustment was then performed prior to acquisition with a ^1H NOESYpresat sequence. 128 free induction decay (FID) transients were collected and summed over a 20 ppm sweep width from -5 to +15 ppm, with a time domain resolution of 32k data points.

After acquisition, spectra were imported into the iNMR (Nucleomatica, Molfetta, Italy) software package for further processing. Before Fourier transform, time domain spectra were zero filled to 64 k data points and multiplied with a 1 Hz exponentially decaying line broadening function. Using iNMR's batch processing, spectra were automatically phased, baseline corrected and referenced to TSP.

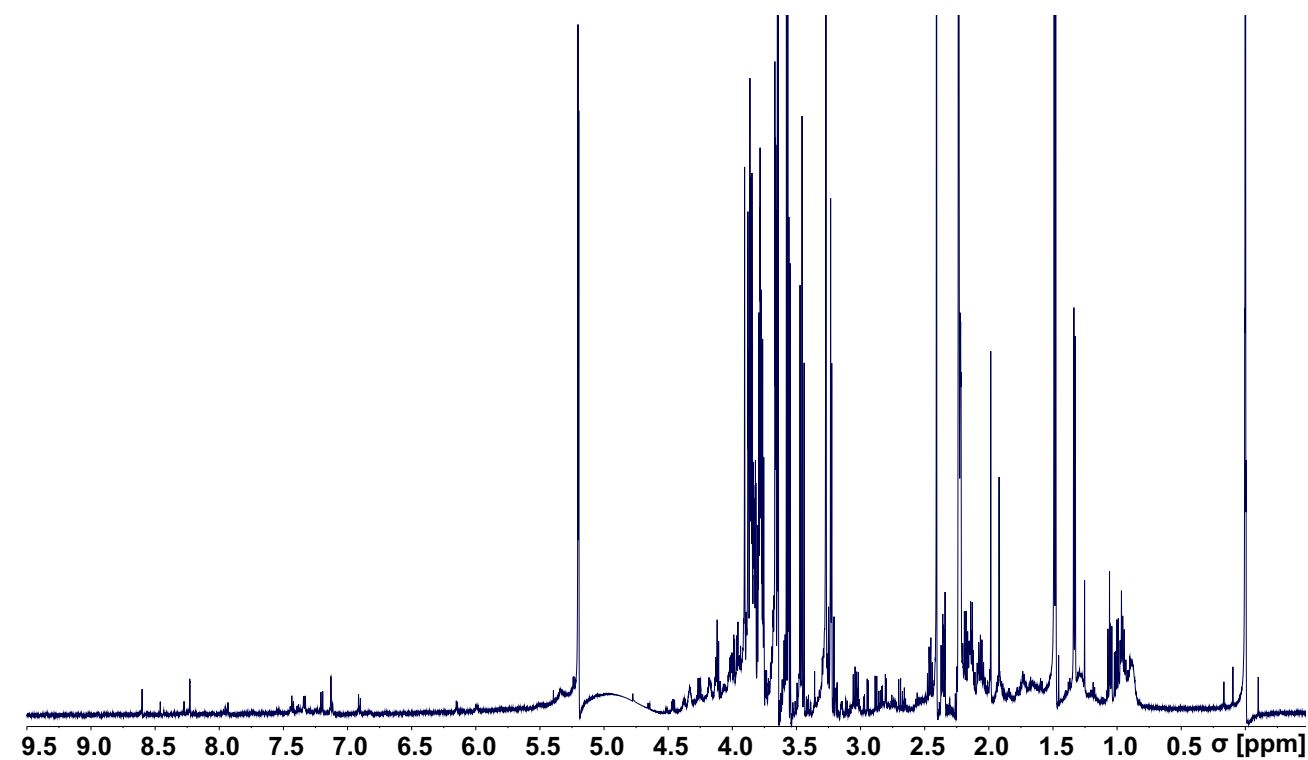
III.3.3. Results and Discussion

An aliquot of the highly concentrated extract from wild type *C. elegans*, grown by liquid culture, produced a very signal rich NMR spectrum (displayed in figure III.7a), when acquired with the tube-NMR based method. The signal overlap in the aliphatic region reveals the problem of complex mixture NMR and highlights the usefulness of a prior separation step for an in-depth metabolome characterisation.

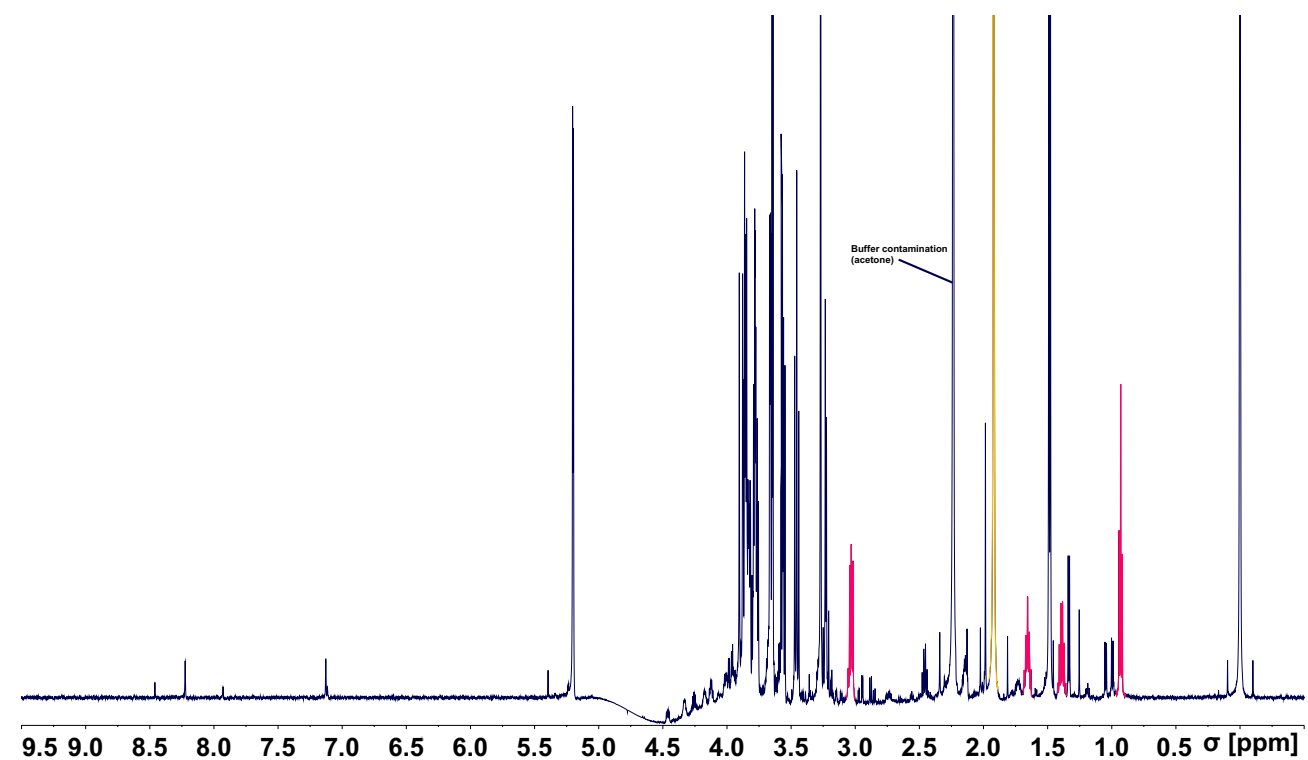
III.3.3.1. The ion pair reagent needs to be deuterated

Before any experiments were performed, it was tested whether the ion pair reagent (tributylamine) needs to be deuterated, when used with the *C. elegans* matrix. To this end, a separation was performed using non-deuterated modifiers and ion pair reagent. The first minute of the eluent (after void volume) was collected, dried and analysed by tube NMR (figure III.7b). Even though the signal of the first LC fraction is usually by far the most intense, the resonances of the ion pair's butyl moieties and acetic acid singlet occupy a considerable share of the chemical shift range. Comparing the neat spectrum with the first fraction, it can be seen that the separation worked as several peaks are missing and therefore must be retained. However, it also becomes apparent that the additive's resonances overlap or even completely mask several metabolite peaks, as the 10 mM of the ion pair is in a typical concentration range of NMR visible metabolites [578]. For this reason, Simpson *et al.*[525], who used tetrabutylammonium ion pairing to separate natural organic matter, primarily interpreted the aromatic region of their acquisition. Even worse, subsequent fractions will contain less total metabolite signal. Therefore the metabolite to additive signal intensity ratio is bound to decrease, but the amplifier gain will still be set to the height of the additives. As the amplifier resolution is limited, this could negatively affect the S/N for lesser intense metabolites; the reason for which solvent suppression is generally used. Simpson *et al.* also used a WET (water suppression enhanced through T1 effects) [451, 529] water suppression sequence adapted to selectively suppress the ion pair's resonances. However, not only is this adding complexity, but it still fails to solve the problem of signal overlap - in fact, it will even attenuate signals close to the suppressed resonances.

For those reasons, fully deuterated chemicals, *i.e.* d_{27} -tributylamine, d_4 -acetic acid, D_2O and d_3 -Acetonitrile, were used for all subsequent experiments.



(a)



(b)

Figure III.7.: Spectral overlap between *C. elegans* extract and ion pair reagent:

Figure III.7a displays the complexity of the *C. elegans* wild-type extract used in this study. The first minute fraction (proceeding the void volume) from a LC separation, with non-deuterated ion pair (red) and acetic acid (orange) is shown in figure III.7b.

III.3.3.2. Online LC-NMR sensitivity is too low for the *C. elegans* extract

As a first test for the complementary application of ion pair and reversed phase chromatography, online LC-NMR was performed. The eluent of the LC was directly coupled to the 60 μ l flow cell probe inside a 600 MHz spectrometer. For reversed phase, a generic, linear 45 minute gradient with initial and final 5 minute hold and for ion pairing an optimised tributylamine gradient according to Luo *et al.*[378] was used. Results are displayed in figure III.8. The deuterated chemicals behaved as anticipated, compared to acquisition with protonated additives (figure III.8c). Despite a small residual artefact from the water suppression, solvent suppression was very effective for the ion pairing run (figure III.8b) and most of the reversed phase method (figure III.8a). The artefact at the end of the RP method is most likely due to absence of D_2O and therefore the failing of the frequency lock. Thus, the software struggles to automatically find and adjust the suppression frequency.

In the reversed phase run, every visible resonance, except for an unidentified doublet and potential artifact at 9 minutes and 2.7 ppm, eluted with the void volume. One singlet, at 38.7 minutes and 2.63 ppm was the only visibly retained compound for the ion pair method. This singlet was identified as succinic acid, that was shifted from its usual position at 2.41. This is due to two reasons. Firstly, the ppm scale is usually referenced (locked) to water (D_2O). However, the chemical shift of a solvent peak inside another solvent depends on their ratio [205, 305]. If now the ratio and thus resonance changes (throughout the gradient), but the water resonance is kept fixed at the centre of the spectrum, analyte resonances will shift (as can be seen with the non deuterated ion pair in figure III.8c). A second and probably more important reason for this ppm shift, is that due to the acetate buffered solvent, one of succinic acid's carboxylic acid sites will be deprotonated and therefore unshielded, compared to an acquisition using a pH 7 phosphate buffer.

As in the neat spectrum succinic acid was the second most intense peak, it is obvious that the ion pairing separation worked, but the subsequent NMR detection did not provide enough sensitivity to detect other compounds from this sample - the number one problem of on-flow LC-NMR [6, 535, 606]. To address this, Spraul *et al.*[535] demonstrated that by operating the detector at cryogenic temperatures, up to a four fold improvement in S/N can be achieved. Even though this was not possible for this study, it remains as a possible future option. Furthermore, another LC-NMR weakness became apparent: the full width half maximum of succinic acid increased from 1 to 6 Hz, when comparing tube based and

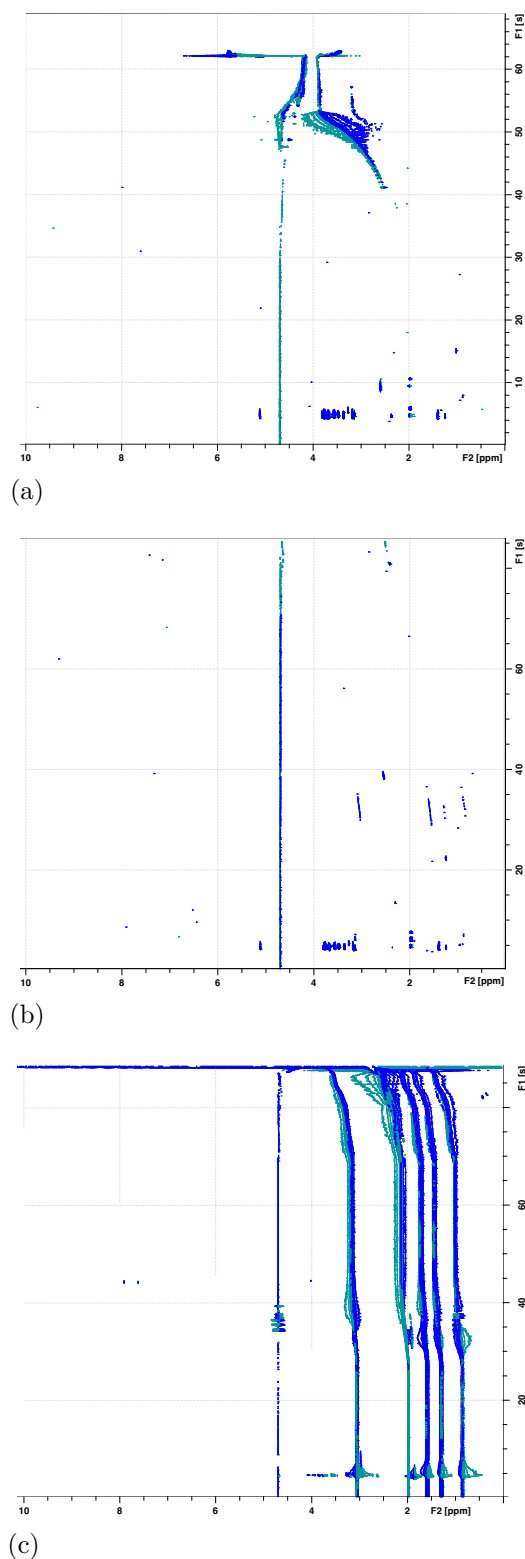


Figure III.8.: **On-flow LC-NMR experiments:**

Figure III.8a and III.8b show an online LC-NMR reversed phase and a tributylamine based ion-pairing separation, respectively. In both cases, solvents, modifier (acetic acid) and the ion pair were fully deuterated. In contrast, figure III.8c displays an ion pair separation, in which ion pair and modifier were not deuterated.

online spectrum. The deterioration of the shims is caused by two phenomena that cannot be addressed readily: The mix of different excitation states in the flow probe as "fresh" nuclei are continuously introduced [6] and a constantly changing field homogeneity by the changing ratios of the two solvents.

As using deuterated ion pair chromatography seemed to work in principle, and more compounds than just succinic acid were retained when the neat extract and the first fraction are compared, it was decided to continue, but change to an offline method. By fractionating the whole run and using conventional tube based NMR to analyse the individual fractions, each sample can be prepared with the same pH and solvent composition, individually shimmed and acquired using a desired number of scans.

III.3.3.3. Offline ion pair chromatography compliments reversed phase

The eluent of both the reversed phase and ion pair chromatography was collected in one minute fractions, dried and analysed by conventional tube NMR. The individual and partially assigned spectra can be seen in figure III.9. As assumed, the bulk of NMR visible material elutes within the first two fractions after the void volume. However, particularly in the aromatic region the ion pairing method appears to retain a fair share of resonances. All retained resonances were assigned as far as possible, using the assignments of the previous chapters, shift databases, computer assisted fitting software and retention times from the publications that characterised the ion pairing method [378, 375].

In the reversed phase spectra (figure III.9a), resonances characteristic for a steroid backbone [324, 36] can be seen at around 20 minutes. Even though this is exciting, as several *C. elegans* specific steroid based signalling molecules have been discovered and also shown to have a role in ageing [427, 192, 355], those molecules are most likely to be found at a much lower concentration. Although more detailed NMR studies would be necessary for confirmation, this compound is most likely cholesterol, as *C. elegans* needs to be supplemented with cholesterol at 1 mg per litre medium, due to its impaired sterol biosynthesis [339]. Furthermore, at an even higher organic solvent ratio, several constituents of lipids, such as alkyl-moities ($\sigma \approx 0.8$ ppm) and alkenes ($\sigma \approx 2.1$ ppm), most likely from unsaturated fatty acids, eluted.

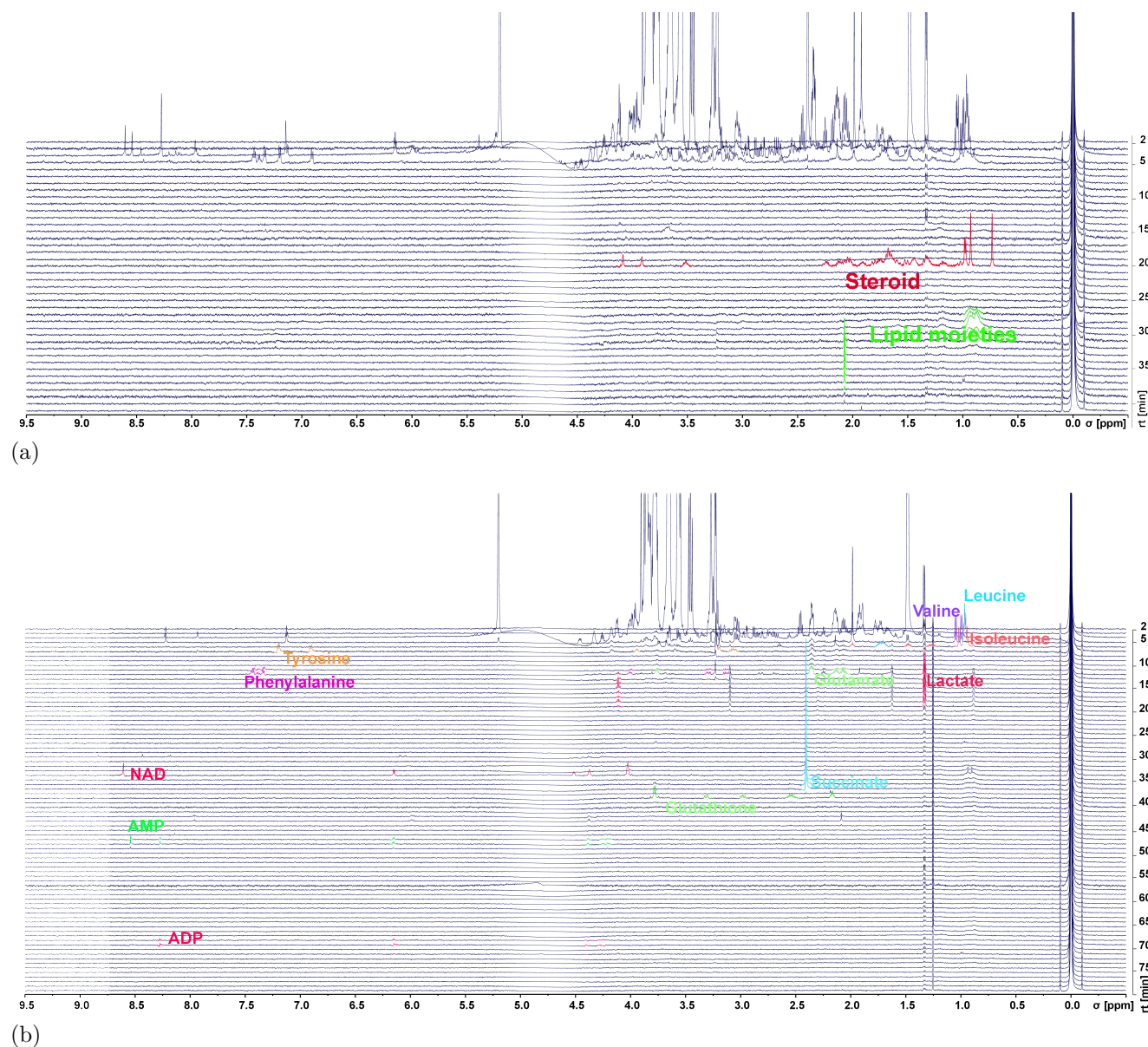


Figure III.9.: **Fractions from offline LC-NMR experiments:**

1 minute LC fractions were collected offline and analysed by conventional tube-based NMR after solvent evaporation and resolubilisation in D_2O buffer. The separation was performed using reversed phase (figure III.9a) and ion pairing (figure III.9b) chromatography. Resonances of an annotation are in the same colour as their respective label.

Compared to reversed phase, in the ion pair separation (figure III.9b) several specific metabolites were assigned. As expected, most of those were molecules attributed to the central carbon metabolism, such as amino acids, small organic acids, nucleotides or redox equivalents. All of those can be observed in 1D ^1H , without separation. However, many metabolites with aromatic resonances cannot routinely be assigned with high confidence by 1D or even 2D NMR. Differentiating between different levels of phosphorylation of the energy or redox equivalents, such as AMP, ADP and ATP and NAD and NADP poses a regular assignment challenge. This is because their diagnostic ribose protons, which are the closest to the differentiating moieties, the phosphoric acid groups, are usually masked by other sugars, due to their relatively low relative abundance. Furthermore, the ppm shift effect of different phosphorylation levels on these protons is not very pronounced and is easily impacted by small pH variation. However, as can be seen in the chromatogram, their retention times vary vastly. For example the 47 minute retention of AMP is extended to 69 minutes by addition of just another phosphoric acid group (ADP).

It must be noted, that there are several other retained metabolites, which could be elucidated readily, if one of their resonances were to be found as biomarker. Particularly, as NMR tube samples can be conveniently stored long term for future analysis. Therefore offline acquired ion pair chromatography fractions with deuterated additives are a helpful addition to the spectroscopist's toolset for characterising unknown matrices.

III.4. Workflow for GC-MS TIC alignment for easy visual interpretation and application to longevity mutants

III.4.1. Background and introduction

As discussed in chapter II.5.4, the current GC-MS workflow, which is explained in the theory and methods chapter (A3.1.4), has a major disadvantage: as it only considers compounds from the Fiehn retention time and fragmentation library [323], it is not truly untargeted and therefore not in the "true" metabolomics spirit [204]. The post-processing step using the GaVin software package [38] requires some manual input, and so can be time consuming and is not well suited for a quick, initial overview. Furthermore, the workflow does not yet include a retention time correction. Even with the small deviations of GC, this may cause problems if two closely eluting compounds share the same quantification ion.

Just like in NMR, full deconvolution and integration of all compounds may not be necessary for an initial assessment of the data. In NMR, spectral overlays, *e.g.* as used for the off-line fractions of the ion pair separation in the last chapter (see figure III.9), provide a quick, untargeted, and powerful shortcut to identify differentially abundant metabolites. Analytes can then be subsequently identified and deconvoluted, by a manual [578] or automatic [617, 229] spectral fitting approach, to yield quantitative data.

However, NMR spectra are often aligned before spectral overlay, to remove small frequency shifts due to pH deviations [598]. Many alignment procedures have also been reported for chromatographic/mass spectral data [332]. Two non-linear algorithms of particular interest are the LOESS regression of xcms (discussed in A3.1.3.1), and the open-source software package "FlagMe" [499]. Even though it may have been initially intended for GCMS alignment, xcms ($x = G/L$) has not been widely used for this purpose. Despite its otherwise excellent performance for high resolution data, it is not able to make use of (unit mass) GC-EI-MS fragmentation patterns. In contrast, FlagMe tries to align AMDIS [539] components (A3.1.4), *i.e.* collected fragmentation patterns, across samples without identifying them first, based solely on their spectral similarity (dot product) [540]. This is done by pairwise comparison of all samples using an algorithm, similar to the ones used in multiple

(gene) sequence alignment [499]. However, it was found that this often introduces abrupt retention time shifts, rather than the "smooth" LOESS alignment models generated by xcms.

Therefore this chapter tries to combine both alignment approaches, by using AMDIS components to LOESS align total ion chromatograms (TIC) of GC-MS files. The developed workflow will then be used to analyse a dataset comparing metabolic differences of *C. elegans* mutant strains.

Strain	Gene(Allele)	Complex	Extension [%]	Literature
N2 wildtype	-	-	-	-
MQ1333	<i>nuo-6</i> (qm200)	I	70	[649]
MQ989	<i>isp-1;ctb-1</i>	III	60	[169]
CW152	<i>gas-1</i> (fc21)	I	-33	[303]
TK1	<i>mev-1</i> (kn1)	II	-35	[274]

Table III.3.: ***C. elegans* strains used to test the GC-MS TIC alignment:**

This table includes the strain name, the knocked down allele and the respiratory chain complex number of the to be analysed mutant strains. Furthermore, the observed life-span extension (compared to N2 wild type) according to literature is given.

All four mutant strains have mutations in subunits of mitochondrial electron transport chain complexes, as outlined in table III.3. Even though all mutations impair mitochondrial function, two of them are short-lived and two long-lived. The molecular links between mitochondrial function and ageing is a long and ongoing scientific debate [26]. However, metabolomics may generate new hypotheses.

For a detailed discussion of mutations, which increase lifespan in *C. elegans*, please refer to chapter V. and in particular section V.1.3.6.

III.4.2. Materials and Methods

III.4.2.1. Sample generation, extraction and acquisition

Worm culture was performed by Dr. Marija Sumakovic, at the lab of Dr. Aleksandra Trifunovic, at the University of Cologne, Germany.

The following number of replicates (plates) per strain were prepared: wild type $n = 11$, MQ989 $n = 13$, CW152 $n = 15$, TK1 $n = 14$ and MQ1333 $n = 14$.

All worms were grown according to standard procedures (see protocol in the appendix B1). Synchronised worms were harvested at young adult stage, before the onset of egg laying. They were snap frozen and transported to Imperial College London, in a frozen state.

Samples were extracted using 80 % methanol and a bead-beater, according to the standard protocol, outlined in the appendix (B2). A QC sample was prepared by pooling 10 % of each sample. The remainder was partitioned into three technical replicates, the solvent removed by a vacuum concentrator and samples stored at -80°C , until analysis.

One of the three technical replicates of the samples was analysed using GC-MS. The two step methoximation with subsequent silylation derivatisation procedure, according to the standard protocol (B4), was performed. For more information, please also refer to the appropriate section in the theory and methods chapter A2.3.3.2.

Samples were prepared and run in a single batch. The samples were run in replicate blocks (*i.e.* replicate one of all samples, replicate two...). QC samples were injected in between to monitor runtime performance. A blank was used at the beginning of the acquisition to assess system contamination. Acquisition parameters were set according to the the Fiehn method [323].

III.4.2.2. Data processing

Data files were converted to the CDF format within Chemstation (Agilent, UK).

III.4.2.2.1. Retention time alignment (for TIC overlay)

Inside AMDIS, a representative QC file was used to build a "model component" library, consisting of unique, empirical fragmentation spectra, with retention indices locked to d_{27} -myristic acid. This library was then selected as the target library, for the deconvolution of all sample set files. The AMDIS output (FIN files) were parsed into R, to yield a "model component" by sample matrix ($C \times S$) of the retention times (with the dimensions $c = 1, 2, \dots, 259$ and $s = 1, 2, \dots, 67$). For each component, its median retention time was calculated, yielding the 259 elements long median retention time vector $\hat{r}$. Correction is performed sample by sample. Each sample's component retention time vector r_s will be aligned to $\hat{r}$, and values in-between interpolated. To this end, the deviation of r_s and $\hat{r}$ is calculated, according to equation III.1.

$$\Delta r_s = \hat{r} - r_s \quad (\text{III.1})$$

The retention time deviation (from the median retention time) Δr_s is only known at $c = 259$ retention times. However, it must be interpolated for each datapoint across the chromatogram. Thus, for each sample s , a LOESS regression model M_s is fitted according to equation III.2; with Δr_s and $\hat{r}$ as independent and dependant variable, respectively. The LOESS regression model used first order polynomials and a smoothing parameter ("span") of 0.6.

$$M_s = \text{loess}(\Delta r_s \sim \hat{r}) \quad (\text{III.2})$$

M_s can then be used to predict a sample's retention time deviation (from the median retention time) at each scantime x_s .

$$\Delta x_s = M_s(x_s) \quad (\text{III.3})$$

According to equation III.3, this yields the scan time deviation vector Δx_s , which gives x_s 's deviation from the median scan time vector $\hat{x}$.

$$x'_s = x_s + \hat{x} \quad (\text{III.4})$$

As shown in equation III.4, the new, aligned scan time vector x'_s , can easily be computed by summing x_s and $\hat{x}$. This procedure is used to compute x'_s for all s samples.

After this, (CDF) sample files, with x'_s as the new scan time vector, were created and a second iteration of the retention time alignment workflow was performed. Finally, TICs were overlaid, by using x''_s as a common x-axis for all samples.

III.4.2.2.2. Standard GC-MS processing workflow

Next to the the TIC alignment approach, the regular GC-MS workflow according to the Fiehn method (see A2.3.3.2) was used (with a database version from 07/2008). As usual, post processing was performed in GaVin. The integrated intensity matrix was subject to probabilistic quotient normalisation. The normalisation factors were also applied to the aligned TIC spectra and the differential features, not detected by the standard workflow.

III.4.2.2.3. Processing of differential features, not found by the standard workflow

Some peaks were identified as potentially different between the mutant strains on the basis of the TIC overlay, but were not identified using the standard workflow. These were followed up in more detail. A combination of running authentic standards, and querying the Golm and NIST databases were used for (putative) identification, prior to quantification. Three additional metabolites could be identified: fructose-6-phosphate, 2-aminoadipic acid and 2-hydroxyglutaric acid. Intensity integration of these was performed by xcms in a targeted manner, using the original, unaligned data files.

To this end, quantitation ions for those three compounds were chosen to be unique, compared to closely eluting components. (6 TMS)-fructose-6-phosphate (F6P) and (6 TMS)-glucose-6-phosphate (G6P) do not have differential ions that could be used for their integration (315 m/z, 357 m/z, 387 m/z,). However, the GC method allows for sufficient separation of the two; F6P elutes at 21.33 minutes, whereas G6P does so at 21.44 minutes. (3 TMS)-2-aminoadipic acid elutes at 15.35 minutes and has unique quantitation ions at 260 m/z, 217 m/z and 128 m/z. (3 TMS)-2-hydroxyglutaric acid (2HG) elutes at 13.76 minutes and coincides with (2 TMS + 1 MeOx)- α -ketoglutaric acid (aKG) (13.78 minutes). However, both have mutually exclusive and abundant fragmentation ions, which can be used to discern between the two and quantitate them (2HG: 129 m/z, 203 m/z, 247m/z,

288 m/z; aKG: 198 m/z, 288 m/z).

To test whether F6P, 2AA and 2HG were truly different in at least one group, one-way ANOVA tests was used and p-values corrected for multiple testing, by controlling the false-discovery rate (FDR) [41]. In a few cases, when deemed necessary to justify fold changes, a metabolite concentration between two groups was compared using a two-tailed (Welch's) t-test.

III.4.3. Results and Discussion

III.4.3.1. Workflow for the GC-MS retention time alignment for TIC overlay

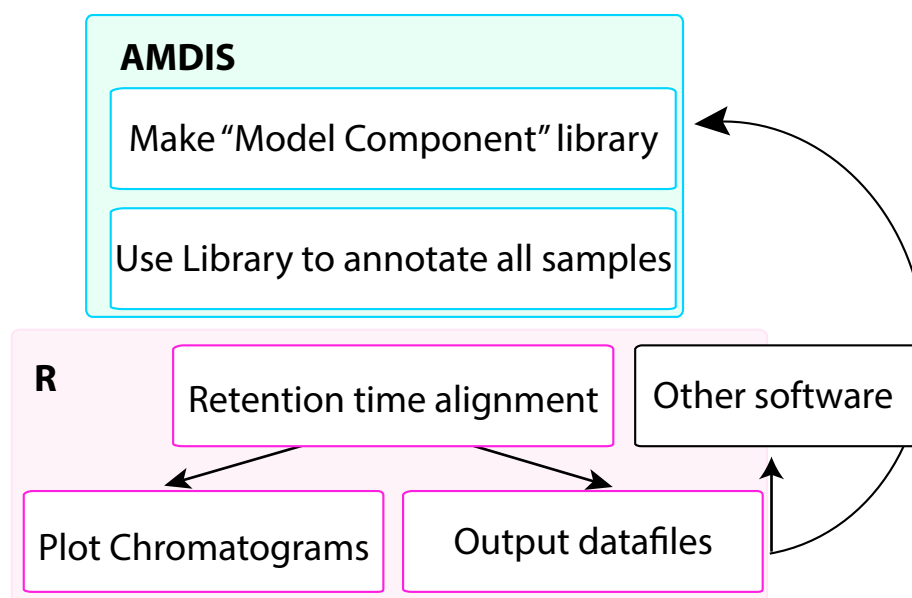


Figure III.10.: Outline of the proposed (TIC) GC-MS alignment workflow

Figure III.10 outlines the workflow for retention time alignment of the GC-MS samples. The first step of retention time alignment is the grouping step (see A3.1.3.1). The same compound needs to be detected across all samples, to gauge the retention time drift from sample to sample. For this step, the FlagMe software package [499] tries to match every component (fragmentation pattern) that AMDIS extracts from each sample against the ones in all other samples, to identify the same compounds. Components are often in the order of several hundred and the algorithm does not discern between "high quality" and "low quality" components. Low quality components do not represent unique fragmentation patterns and therefore are often mismatched to the wrong components inside the other samples. In the workflow proposed here, the components used for grouping and thus retention time alignment are defined by the user. This can be done conveniently inside AMDIS, by adding individual components of a sample to a new library. Ideally, those components are present in the majority of the samples and have fragmentation patterns distinct from closely eluting neighbours. A pooled QC sample is well suited for this purpose.

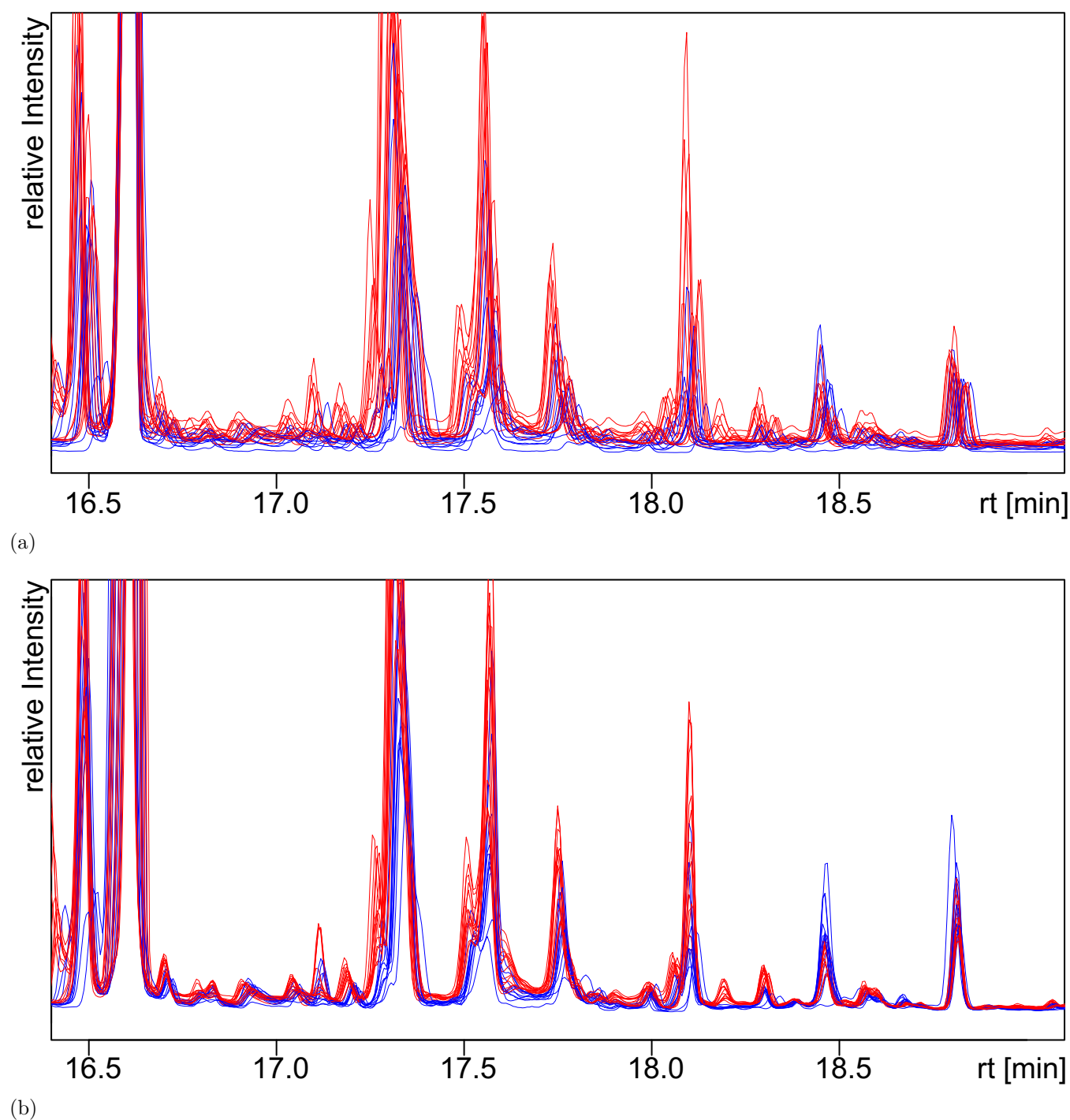


Figure III.11.: **TIC before and after retention time alignment:**

Excerpt of the normalised total ion chromatograms (TIC) before (figure III.11a) and after (figure III.11b) retention time alignment. Wild type is blue, MEV-1 is red.

In the second step, the empirically built "model component" library is used to identify the chosen "model components" across all samples. Again, this can conveniently be performed with AMDIS. A LOESS regression model, built for each sample, from the retention times of the "model components" is then used to align the retention times to the median retention time (see A3.1.3.1). After alignment, the total ion chromatograms (TIC) can be plotted. Alternatively, the aligned data can be output into generic mass spectral data formats (netCDF), for another alignment iteration, or processing using another software suite.

Figure III.11 shows a retention time region of two sample groups, before (figure III.11a) and after (figure III.11b) retention time alignment. Alignment clearly improves the interpretability of the spectral overlay of the different samples, by removing small retention time drifts. In fact, differences in metabolite concentrations of different sample groups can readily be assessed, without the need of mass spectral deconvolution. This represents a quick method for initial data exploration, without restriction to known compounds.

However, alignment procedures should be used with care. Liebeke *et al.* recently showed that alignment of ^1H NMR spectra may not only introduce artefacts, but can also align the wrong peaks, particularly in cases of peaks crossing their positions [365]. Compared to 1D NMR alignment, which is simply based on the similarity of peak shapes [598, 515], the procedure used here is based on (unique) fragmentation patterns. Thus, the chance of mismatches is greatly reduced. Peak crossings are not currently modelled by any GC/LCMS alignment procedure [446], but will be far rarer for GC than NMR (where different peaks will have individual responses to sample factors such as ionic strength). Therefore, whenever an alignment method is used, it is important to follow up with a subsequent deconvolution and quantification of the differential peaks, using unique quantitation ions. However, as Liebeke *et al.* conclude, (^1H NMR) alignment based overlay methods is still a valuable tool for initial data exploration, particularly when unidentified metabolites are present. A similar conclusion may be drawn in the case of TIC GC-MS alignment.

III.4.3.2. TIC alignment generates interesting hypotheses for the *C. elegans* mitochondrial mutant dataset

In collaboration with the University of Cologne, metabolomics was used to analyse four *C. elegans* strains with mutations in the mitochondrial respiratory chain, which either lead to

longevity or premature ageing. To generate hypotheses about underlying mechanisms, it was decided to use GC-MS, as evidence exists that those mutations primarily affect central carbon metabolism, and tricarboxylic acid (TCA) cycle activity in particular [77, 483]. All samples were analysed using the standard deconvolution based workflow, as well as the TIC overlay based approach, proposed above.

Gauging from the aligned and normalised TIC spectra, displayed in the appendix (figure C2.1), there are extensive differences in metabolite concentrations between the different mutants. Using the standard workflow, many of them could be identified and shown to be considerably different from wild type. In agreement with literature [77, 483], these included TCA intermediates and related metabolites, such as succinate, citrate, lactate, glycerol-1-phosphate, as well as amino acids. The dipeptide glycine-proline was found to be increased 1.8 fold ($p < 0.0018$) in the MEV-1 mutant compared to wild type. The same observation was already made by Butler *et al.* [77], who analysed the exudate of long and short lived *C. elegans* mitochondrial mutant strains. This could possible hint at a breakdown of collagen.

Many more peaks were present than identified by the standard workflow (using the Fiehn library), and a considerable fraction of these had concentration differences between groups. Two of the three additional metabolites that could be identified were significantly altered in at least one strain: fructose-6-phosphate (F6P) ($p < 0.651$), 2-aminoadipic acid (2AA) ($p < 0.022$) and 2-hydroxyglutaric acid (2HG) ($p < 0.003$).

Even though F6P was not flagged significantly different by ANOVA, individual two-way comparisons of mutants against wild type (t-test) indicated that GAS-1 was different ($p < 0.0038$). Overall, fructose-6-phosphate showed the same concentration pattern as glucose-6-phosphate ($R^2 = 0.972\%$). This is unsurprising, as both are in a chemical equilibrium, linked by the phosphoglucose isomerase enzyme, and are adjacent in the same pathway (early in glycolysis). It provides a good example of how additional metabolites can be identified, beyond relying on a (widely-used) commercial database. Furthermore, it is an example where both retention time and mass spectra are similar enough that it would likely confound assignment of the library metabolite.

2-aminoadipic acid appears only as an intermediate in the lysine biosynthesis/biodegradation pathway. 2AA is the stable oxidation product of 2-aminodipate-semialdehyde. 2-aminodipate-

semialdehyde itself is routinely used as a biomarker for quantifying oxidative protein damage, as it is a prominent carbonylated (oxidised) protein-amino acid [490, 584, 538, 161]. Recently, proteinaceous 2AA was also proposed as a biomarker for ageing of human skin, where it was found to accumulate inside proteins with age [520]. The cuticle of *C. elegans* and human skin are both comprised of collagen [114, 127, 331]. Thus, it is likely that free 2AA is a biomarker for oxidative protein damage/stress also in *C. elegans* with clear utility for ageing research. In a ^1H NMR profiling study, conducted in collaboration with Sukaina Zeitoun-Ghandour and Stephen Střrzenbaum from King's College London, 2AA was found to be increased during exposure to the oxidative stress inducing chemical N,NŌ-dimethyl-4,4Ō-bipyridinium dichloride (Paraquat) [660]. However, this does not match with the fact that for the two strains in which 2AA was decreased compared to wild type (2.2 fold in NUO-6 ($p < 5 \times 10^{-04}$); 5.7 fold in GAS-1 ($p < 3 \times 10^{-05}$)), increased protein carbonyls are reported [483]. Clearly, 2AA has a role in *C. elegans* ageing and oxidative stress responses, but its full biochemical role will require further investigation.

2-hydroxyglutaric acid is an even more intriguing assignment. Recently, 2HG was shown to be an "onco-metabolite", as it is associated with various forms of cancer [96, 372, 652]. It is found in increased concentrations inside gliomas [96]. In cell models, however, its tumour-inducing effects are reversible upon removal [372]. These effects are due to its ability to competitively inhibit multiple α -ketoglutarate-dependent dioxygenases, as recently shown in *C. elegans*[646]. Some of the affected dioxygenases are histone demethylases, through which 2HG can exert epigenetic effects [374]. It was also shown that 2HG directly affects hypoxia inducible factor HIF-1 by controlling the proline hydroxylase EGL-9 [646], required for HIF-1 hydroxylation and degradation [152]. (R)-2KG increases EGL-9 activity, and (S)-2KG decreases EGL-9 activity [333]. Unfortunately, the GC-MS method used here cannot differentiate between the enantiomers. However, the possibility of different HIF-1 modulation in *C. elegans* mitochondrial mutants is highly interesting. HIF-1 has emerged as a key player [268] and even a requirement [352] for mitochondrial mediated longevity. Recently, it was shown that impaired mitochondrial function in *C. elegans* neurons is enough to induce the mitochondrial unfolded protein response (UPR^{mt}) in the whole animal. UPR^{mt} is required, if not the reason for the longevity mechanism [147]. Thus, a signalling molecule which cell-non-autonomously induces this response must exist [637]. 2-HG, as a modulator of HIF-1, would be an intriguing candidate for this signal. Even more so, as the short-lived GAS-1 mutant and the long-lived NUO-6 mutant both

showed a 12 fold ($p < 0.0005$) and 9 fold ($p < 0.0007$) (median) decrease of 2HG compared to wild type, respectively. Both mutants were also shown to have increased HIF-1 activity and an induced UPR^{mt} [352, 483]. Mitochondrial longevity can also be modulated by RNAi, rather than KO mutations (used in this study). However, silencing must occur in a defined developmental window, between the third and fourth larval stage [137, 487, 147]: RNAi treatment during adulthood has no effect on lifespan extension. Thus, it has been suggested that this lifespan extension occurs via an epigenetic control mechanism [147]. Indeed, RNAi and KO induced mitochondrial longevity appear to be regulated over two distinct mechanisms [649]. They may even lead to additive lifespan extension [649]. This is also intriguing, considering the two proposed modes of action of 2-KG [652]: one mediated over EGL-9 and HIF-1, the other over epigenetic mechanisms. Taken together, 2KG may be a key molecule controlling lifespan regulation, mediated by mitochondrial function.

As well as 2AA and 2HG, several other unidentified peaks were also differentially expressed. However, even without assigning more peaks, the current data serve as an example of how truly untargeted approaches, such as the TIC alignment method presented, can generate novel and interesting hypotheses, which can then be tested using standard life sciences techniques.

III.5. Conclusion

This chapter presented three approaches to help increase metabolite coverage and ease of annotation for NMR spectroscopy and GC-MS. All methods presented in this chapter are readily transferable to other matrices. They may have to be adapted, however; *e.g.* isotope enrichment may not be possible for matrices such as urine or plasma. Still, they can often be obtained in larger volumes and concentrated, to improve their S/N ratio before using the automatic database matching workflow of section III.2.

The next chapter (IV.) will develop and test a method to increase the coverage and ease of annotation for LC-MS based metabolite profiling. This method will then be applied to a large scale *C. elegans* longevity mutant screen in chapter V.

IV.. A tool for data reduction and annotation of LC-MS features

Aims and Ojectives:

- Investigate necessary atom-count constraint for unambiguous formula generation
- Develop an automated workflow for stable isotope labelling based data reduction and formula generation
- Test the developed algorithm on different LC-MS datasets

IV.1. Introduction

IV.1.1. Motivation and Background

The last chapter presented approaches to improve the metabolite coverage and ease of annotation for NMR and GC-MS based profiling platforms. This chapter aims to do the same for LC-MS.

As noted in chapter II., LC-MS excels by its wide coverage and high sensitivity [67]. Although GC-MS displays a similar sensitivity, it requires metabolites to be volatile (after derivatization). Therefore it can rarely be used to analyse compounds larger than 300 Da. NMR spectroscopy, in contrast, has no restrictions on molecular weight. Indeed, the LC-NMR experiments (in section III.3) showed lipid and steroid resonances. However, NMR's lack of sensitivity did not allow a detailed characterisation. LC-MS on the other hand is a powerful tool for lipid analysis [559, 486, 396, 496]. As a large amount of the features detected in chapter II. originate from different lipid species, and lipids are key molecular species in (*C. elegans*) ageing [350, 608, 527, 438, 242, 265, 489], LC-MS is an appropriate choice to explore metabolism and longevity. Furthermore, its sensitivity allows multiple injections from one Petri dish worth of *C. elegans* biomass (3+ plates required for NMR). This enables higher sample throughput, such as required for the longevity mutant screen in chapter V.

IV.1.1.1. Data reduction in untargeted LC-MS profiling

The extreme sensitivity and coverage of current mass spectrometers results in a wealth of information. As much as this opens new opportunities, it introduces major challenges. Compared to the few thousand data points of a NMR spectrum, or the few hundred thousand data points of a unit mass GC-MS run, a single, centroided, high resolution LC-MS run (such as acquired for this study) can have up to 100 million detector count events. This puts a heavy strain on data processing pipelines, especially considering a typical profiling experiment can have up to hundreds of samples [245, 289, 141]. Aside from the computational requirements, efficient signal reduction methods are needed to break down the information into a format interpretable by humans. For the first stage of post-acquisition signal reduction, LC-MS processing tools [531, 302, 547, 482] contain sophisticated filtering algorithms [301, 180] and methods for peak detection [123, 610, 566]. This step greatly

reduces the amount of "noise", but one is still left with several tens or even hundreds of thousands of detected peaks, usually referred to as features. "Features" is an umbrella term for chemical compounds, their isotope peaks, adducts and clusters, and the isotopes thereof. Combinations of rule and correlation based approaches have been proposed to further reduce this potentially redundant information [338, 64].

Afterwards, one must still differentiate between a metabolite originating from the biological system of interest and other "undesired" chemical entities. It is becoming more and more apparent that unwanted contaminants, such as plasticisers or polymers [574, 216], make up a sizeable proportion of all observable peaks [31, 196, 195, 194]. Common sources of contamination include solvents [629, 306], storage containers [645, 663, 546], ambient air [517] and even parts of instrumentation [597]. They may also be introduced by sample handling procedures such as autoclaving [418] or the use of certain gloves during sample preparation [546]. A common way for removing contaminating signal post-acquisition is the comparison of blanks [394]. However, this is not always feasible, particularly when retention time alignment becomes necessary [531]. Another, very effective way of differentiating between biological compound and introduced contaminant is by using stable isotopes. If an organism can be grown on stable isotope enriched medium, a change in the isotopic pattern in a mass spectral peak indicates it must be of biological origin. The effectiveness of this approach to prize out genuine metabolite signals has already been demonstrated for algae [31], bacteria [29] and *Arabidopsis thaliana* [196, 195, 194].

IV.1.1.2. Determination of a feature's molecular formula

Another challenge encountered with LC-MS based profiling is compound annotation. Even though *de novo* elucidation from electron impact (EI) GC-MS spectra can be challenging, annotation of known compounds is usually straightforward. This is due to exhaustive open source [335] and commercially available [25, 323] retention time locked fragmentation databases. Similar databases exist for LC-MS, but are still far from similar coverage [530, 251, 630]. Furthermore, retention times are generally less reproducible than for GC, and fragment patterns may be less characteristic of individual metabolites. However, with recent advances in technology, accurate mass measurements close to sub-ppm accuracy have become possible. This performance used to be the domain of (specialized and expensive) Fourier transform ion cyclotron resonance (FT-ICR) instrumentation, but is now

more widely available for routine profiling applications through ToF-MS instrumentation. High mass accuracy, particularly in conjunction with isotope patterns, vastly improves database search accuracy [319]. It may even allow an *ab initio* molecular formula calculation for low weight molecules, especially with constraints of the expected formula space, for example for lipids only [501]. However, these calculations generally fail to give unique empirical formulae for larger molecules, particularly without additional constraints of the formula space [319, 320].

Rodgers *et al.* [501], have proposed a method by which the theoretical possible formula space can be constrained with the help of stable isotope labelling. They compared metabolites of ^{13}C enriched bacterial extracts with their native ^{12}C counterparts, which, like isotopically labelled quantification standards [639], elute at the same retention time. If the compound is fully labelled (and singly charged), the integer mass difference between the heavy and light version of the metabolite equals its number of carbon atoms. Rodgers *et al.* hypothesise that the knowledge of the exact number of carbon atoms and accurate mass at least triples the upper mass limit for calculating the elemental composition.

Hegeman *et al.* [234] extended that approach, by constraining both carbon ($^{12}\text{C}/^{13}\text{C}$) and nitrogen ($^{14}\text{N}/^{15}\text{N}$) atoms in the model plant *Arabidopsis thaliana*. By means of a simulation, they estimated that this double constraint and 3 ppm mass accuracy can "consistently provide unique assignments up to 500-600 amu". In their experiment, they were able to generate unique assignments in 87% of the cases, compared to 20% without constraint. Bowen *et al.* [58], showed that using C+N double isotope constraint was more powerful for molecular formula calculation than the use of isotope abundance patterns [319], particularly at masses above 500 Da. Giavlisco *et al.* also constrained the carbon atoms of *Arabidopsis thaliana* to assign metabolites with the help of databases, such as pubchem [56] and KNApSAcK [2]. They used both direct infusion FTICR [195] and LC-FTICR [196] for their acquisitions. In 2011 they extended their LC-MS *A. thaliana* constraint approach to a $^{13}\text{C}/^{15}\text{N}/^{34}\text{S}$ regime [194]. This allowed them to consider each peak as one of three possible common adducts, instead of relying on it to be the $H^{+/-}$ adduct, like during their previous studies. Baran *et al.* [31] presented a manual workflow to annotate metabolites from the cyanobacterium *Synechococcus sp. PCC 7002* by carbon constraints. They also performed a subsequent verification of their hits by MS/MS fragmentation studies. However, a large subset of their formulae (generated by a vendor's proprietary software tool)

was not found in any database. Therefore, *ab initio* calculation of elemental composition should be preferred over sole database matching, if possible.

However, a general theoretical analysis on the power and limitations of different atom-count constraints is still lacking, as well as an automated workflow which can convert the information from an LC-MS dataset. These points, which are the basis for this chapter, are outlined in the next section.

IV.1.2. This experiment outline

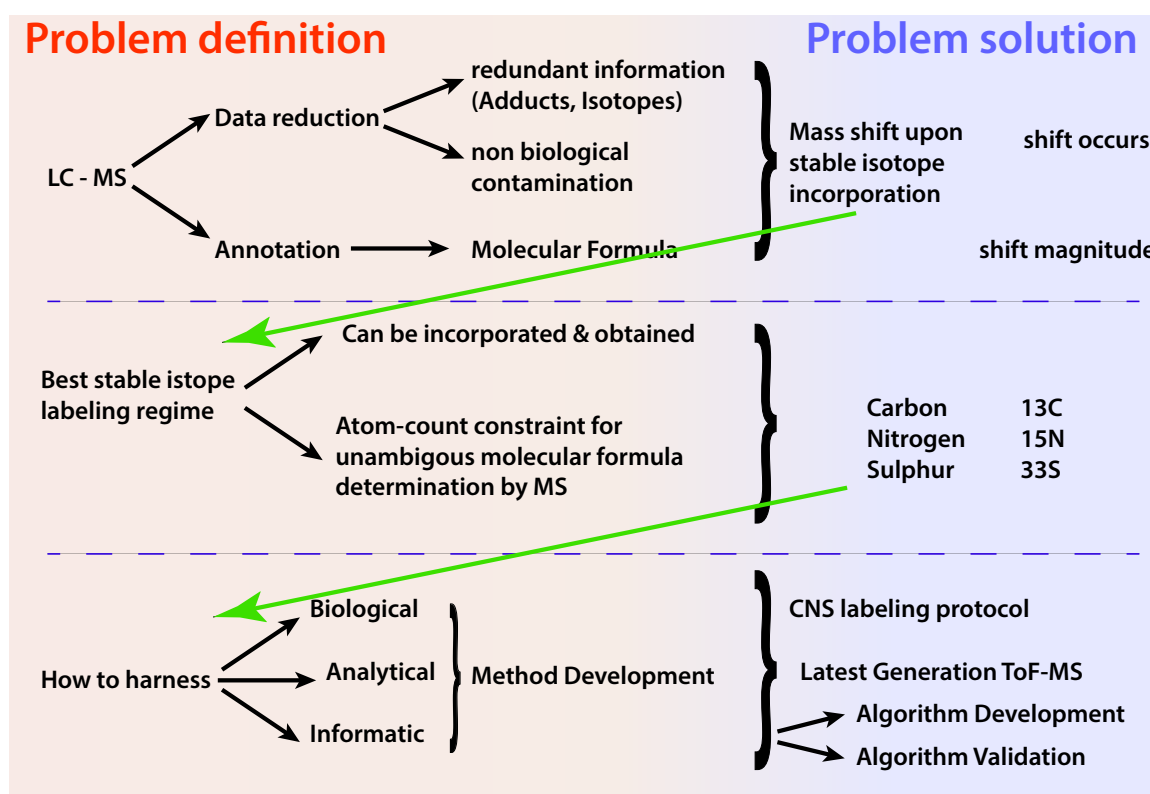


Figure IV.1.: Outline of the chapter

This chapter, which is outlined in figure IV.1, is divided into three major parts. 1. A theoretical and simulation-based test; 2. the development of a processing algorithm; and 3. the testing of this algorithm on four LC-MS datasets.

The simulation uses databases of molecular compositions. This section will investigate what m/z instrument accuracy (ppm) is required to yield unambiguous results when a measured mass is queried against a database, or when an *ab initio* molecular formula calculation of this mass is performed. It will also explore the impact of different atom-count constraints, in order to find a combination which allows unambiguous assignment of all molecules up to 1 kDa at a typical modern ToF mass analyser instrument accuracy. The dependency between the molecular weight of a compound and its required instrument resolution for unambiguous assignment will also be discussed.

In the second part of this chapter, an automated workflow that uses stable isotope enriched samples to reduce LC-MS datasets to non-redundant, biologically relevant information will be developed. Furthermore, the software package also generates molecular formulae based on the exact mass and isotope shift information of biologically relevant features. To apply this algorithm in the third section, a protocol for introducing carbon, nitrogen and sulphur stable isotopes into *E. coli* and *C. elegans* is developed.

In a final section, this workflow is tested using LC-MS acquisitions of carbon, nitrogen and sulphur stable isotope enriched *E. coli* extracts. The sample set consists of two chromatography modes, hydrophilic liquid interaction (HILIC) and reversed phase, each in positive and negative ionisation mode.¹

¹Even though the same four conditions were acquired for isotope enriched *C. elegans* samples, they were not used. This was due to the use of an early stage protocol during samples preparation, which led to incomplete isotopic labelling. As the algorithm is not ideal for working with samples of mixed isotopic labelling levels, it was decided to refrain from an automated analysis of these samples. The problem will be discussed in the results section. A new, fully labelled batch of isotope enriched *C. elegans* samples was used for data reduction and formula annotation in chapter V.

IV.2. Methods

IV.2.1. Sample generation and metabolite extraction

Standard worm culture protocols are not suitable for isotopic labelling experiments. As discussed in section IV.3.2 a protocol that allows introduction of stable isotope labels had to be developed, including determining the minimum sulphur requirement for *E. coli*. The metabolite extraction was performed according to the consensus protocol, developed in the extraction comparison chapter (II.).

IV.2.1.1. Preparation of a modular base minimal medium

Gutnick minimal medium [218] was used as a basis for labelling experiments, with the sulphur-containing compounds (e.g. sulphate mineral salts) replaced with sulphur-free alternatives. This resulted in the following modified protocol:

Salt	Amount	g/L	mg/L
H_3BO_3	2.85	X	
$MnCl_2 \times 4 H_2O$	1.85	X	
$FeCl_2$	0.62	X	
Na_2 Tartrate $\times 2H_2O$	1.77	X	X
$CuCl_2 \times 2H_2O$	26.5		X
$ZnCl_2 \times 2H_2O$	20.0		X
$CoCl_2 \times 6H_2O$	46.4		X
$Na_2MoO_4 \times 2H_2O$	25.2		X
Na_2SeO_3	20.0		X
Na_4 EDTA	0.5	X	

Table IV.1.: **Trace element solution (1000x) for modified Gutnick medium**

Sterile filtered before aliquoting and storage at -40° C.

For 100 ml of medium, 20.0 ml of 5x base medium (table IV.2) and 100 μ l trace element solution (table IV.1) are combined. The remaining 80 ml can be taken from natural abundance or stable isotope labelled glucose, ammonium chloride and sodium sulphate stock solutions and topped up with water to 100 ml, if needed. For the NH_4Cl , 10 ml of a 100 mM (10 mM final) solution and for the glucose, 10 ml of a 4 % stock were used. For the sulphate, 4.5 mg Na_2SO_4 , dissolved in 10 ml was added. After sterile filtering the base/trace element mix with the carbon/nitrogen/sulphur source, 50 ml of autoclaved

water was added. The prepared growth media were kept at 4° C prior to use, either in autoclaved glass bottles or sterile disposable plasticware. To minimise non-biological contaminants, only ultra purity grade water was used.

IV.2.1.2. Determining the sulphur requirement of *E. Coli* on the modified minimal medium

Salt	Amount [g/L]
K_2SCl_2	4.3
K_2HPO_4	67.5
KH_2PO_4	23.5
$MgCl_2 \times 6H_2O$	0.4

Table IV.2.: **Base medium (5x):**
Autoclave before storage.

To minimise the consumption of expensive ^{33}S , the minimum sulphur requirement for the *E. coli* K-12 wild type strain NCM3722 on the modified minimum medium was determined. *E. coli* growth curves were recorded for Na_2SO_4 concentrations corresponding to 1, 5, 20 and 185 mg/L sulphur in the final medium. Growth was performed under the same conditions used for actual bacterial culture for worm labelling: 50 ml of growth medium were incubated in a 250 ml flasks (325 rpm, 36° C). The flasks

were sealed with an autoclaved cotton plug to allow gas exchange. Before performing the growth experiments, a starter culture of *E. coli* was grown, to allow the strain to adapt to the minimal medium. 1 ml of the starter culture was pelleted, washed with M9 buffer to remove all residual sulphur, and resuspended in 1 ml M9. A 100 μ l aliquot was used for inoculation. Every 30 min a 0.5 ml aliquot was taken and the OD_{600} measured. M9 was used to dilute the sample if the OD was too high for the spectrophotometer.

IV.2.1.3. Generation of stable isotope labelled *E. coli* samples

The growth medium combinations (60 ml each) in table IV.3 were prepared, using the base minimal medium described in section IV.2.1.1. In order to capture the most diverse metabolome, a 25 ml sample was taken during exponential growth, and the remaining 25 ml were sampled during stationary phase. The labeled cultures were incubated as described above. Before growing the final labeled culture, the *E. coli* NCM3722 strain was grown twice in 5 ml overnight cultures of the corresponding labelled medium, to allow the strain to adapt to the medium and to minimise carry over of unlabelled biomass. The inocula were washed with M9 to remove any residual old medium.

Combination	^{12}C	^{14}N	^{32}S	^{13}C	^{15}N	^{33}S
None	X	X	X			
C		X	X	X		
N	X		X		X	
S	X	X				X
CN			X	X	X	
CNS				X	X	X

Table IV.3.: **Stable isotope label combinations**

Both samples (*i.e.* exponential and stationary phase) were immediately centrifuged (6,000 g, 5 min, 4 °C) and washed with 5 ml of ice cold M9. After resuspension in 1 ml ultra purity water, samples were immediately snap frozen inside 2 ml bead-beater extraction tubes, prefilled with 200 μl of 0.1 mm glass beads. Samples were stored at -80° C until extraction.

IV.2.1.4. Generation of stable isotope labelled *C. elegans* samples

A protocol for producing labelled *C. elegans* samples was also developed. To introduce stable isotopes into *C. elegans*, its food source, *E. coli*, must be labelled. This was done as described in section IV.2.1.3 above, for all possible labelling combinations. The biomass of 100 ml stationary phase *E. coli* culture was sufficient to support the growth of 3-4 full plates of worms to adulthood at 20°C incubation temperature. Before feeding the *E. coli* to the worms, the pellets were washed with M9 buffer to avoid carryover of any growth medium components. The biomass was then suspended in a small residual of M9 and transferred onto the agar plates for *C. elegans* culture. In order to conserve the expensive isotope-enriched biomass, food was not provided in excess at the beginning, but topped up on a daily basis, ensuring worms were always fed *ad libidum*. Worms were sampled once they nearly exceeded all of the bacterial lawn, to minimise bacterial contamination of the samples. Pellets were stored at 4°C until use, but not longer than a week.

Preparation of agar plates was also modified compared to the standard protocol. First, the 1M Mg_2SO_4 solution (1 ml per litre of medium) was exchanged for an equimolar Mg_2Cl_2 solution. Molecular biology grade agarose was used rather than standard bacteriological agar, to reduce non-biological contamination. Most importantly, when introducing the cholesterol, ethanol must be avoided. The first batch of ^{13}C labelled samples was contaminated with ^{12}C from the ethanol. In the second batch, analysed in chapter V., cholesterol was successfully introduced without the use of ethanol. This was done by adding the correct amount directly to the hot agar, followed by dispersion through manual shaking and

10 min sonication inside a 50° C water bath. Plates were stored (for a maximum of two weeks) at 4° C until use.

To ensure complete isotope labelling, the worms were grown over two generations before sampling. An age synchronisation (bleaching) step was performed between the two generations, to ensure only second generation worms were sampled. Sampling was then performed according to the standard protocol.

In all steps, where water was needed, ultra purity grade water was used.

IV.2.1.5. Metabolite extraction

For both *E. coli* and *C. elegans*, metabolite extraction was performed as outlined in B2, using 80 % methanol and a bead-beater.

The growth and stationary phase *E. coli* samples were pooled. Three equal aliquots were prepared of each sample.

IV.2.2. LC-MS data acquisition

Analysis of the samples was performed at Waters Corporation in Manchester, UK, with the help of Matthew Lewis (Imperial College London), Dr Martin Palmer (Waters) and Dr Steven Pringle (Waters). This was in order to analyse the samples on a modern (2012) ToF instrument (Waters Synapt G2-S).

Samples were reconstituted with 250 μ l water and 1:1 water:acetonitrile for reversed phase and HILIC chromatography, respectively. The samples were solubilised by sonication for 10 min and vortexing. Insoluble debris was then removed by centrifugation (10 min, 10,000 g), and the supernatant transferred into silanized high recovery vials. Samples were kept in a Waters Acquity sample manager at 4°C until full loop injection of 5 μ l (without needle overfill) by an I-class Acquity UPLC system.

Reversed phase chromatography was performed using a 2.1 $\times$ 150 mm Waters ACQUITY UPLC HSS T3 column, with 1.7 μ m C_{18} -alkyl functionalised particles. The column was kept at 40° C. A gradient from solvent A (water) to B (acetonitrile), both with 0.1% formic

added, was run (table IV.4). For both, RP and HILIC, the flow rate was 0.5 ml per min.

Time	A[%]	B[%]
0.00	100.0	0.0
0.10	100.0	0.0
15.10	0.0	100.0
15.60	0.0	100.0
15.70	100.0	0.0
19.20	100.0	0.0

Table IV.4.: **Gradient of reversed phase acquisition**

for mixing. This was done to prevent precipitation of any salts due to the high organic content.

The LC eluent was introduced into a Waters Synapt G2-S Q-ToF mass spectrometer. The eluent was evaporated and molecules ionised by an electrospray source heated to 120°C, with desolvation temperature at 550°C and a desolvation gas flow of 800 l/h. The capillary voltage was set to 0.5 kV, for both positive and negative ionisation modes. Sample cone voltage was 25 V and 35 V in positive and negative mode, respectively. 6 mass spectra per second were acquired over a range of 50 to 1200 dal-

tons and a scan time of 0.1 s, with 0.05 interscan delay. The ToF analyser was operated in high sensitivity mode with around 10 thousand resolution, saving the data in centroid mode. To obtain high accuracy mass measurements, the mass spectrometer was calibrated to sodium formate and a lock mass was introduced into the source throughout the run. The lock mass was 200 pg/ μ l Leucine-Enkephaline in a 1:1 mixture of acetonitrile:water. It was infused at a rate of 10 μ l/min and recorded every 30 scans; averaging four lock mass scans each time. The whole system was controlled by MassLynx version 4.1.

HILIC chromatography used a 2.1×150 mm Waters ACQUITY UPLC BEH Amide column with 1.7 μ m particle size was used. Column temperature was 50° C. The gradient, from solvent A to B, is shown in table IV.5. Solvent A was 19:1 (95 %) acetonitrile:water, and solvent B was 1:1 acetonitrile:water. Both contained 10 mM ammonium acetate and 0.04 % NH_4OH . The pH was adjusted to pH 9.0. Solvent A was prepared by mixing the ammonium salt and solution with the water. The acetonitrile was slowly added, while using sonication

Time	A[%]	B[%]
0.00	100.0	0.0
0.10	100.0	0.0
15.10	0.0	100.0
20.0	0.0	100.0
20.1	100.0	0.0
30.0	100.0	0.0

Table IV.5.: **Gradient of HILIC acquisition**

IV.2.3. LC-MS data processing

After acquisition, data were converted into mzXML format and peaks detected using xcms (version 1.33.14). Parameter settings were kept liberal, to allow for the highest coverage possible. The "centWave" algorithm was used with an expected chromatographic peak width between 1 and 10 s and a signal to noise threshold of one. All other parameters were left at default values.

IV.2.4. Programming and simulations

The simulation engine for the theoretical experiments and the algorithm for the data reduction and formula annotation were both implemented in R. However, all parts which required extensive computation were outsourced into multicore C-code. Due to the sheer amount of computational power and space required, most simulations were performed on the Amazon EC2 cloud computing service, running Ubuntu Linux 10.04 LTE. Peak detection for the workflow test datasets as well as the parameter optimisations were also performed on the EC2 in a parallelised fashion. However, the final workflow algorithm was developed to be run on desktop computers. The recommended specifications depend on the number of detected peaks, and retention time and mass error windows. For a good performance, 8 GB RAM are recommended, 16 GB ideal. The code has been tested on Mac OS X 10.6-10.8 and Ubuntu Linux 10 with R 2.13 up to R 3.0.

The PubChem library was downloaded on the 2nd of August 2011, and the EcoCyc library on the 8th of June 2011. Both were parsed using scripts developed in R. Only compositions consisting of the elements CHNOPS with at least one carbon atom were used for the simulations. It was decided to perform all simulations with charge neutral compounds. In cases where a charge was present, for example an acid, the charge was *in silico* "titrated", adding or removing protons. The mass range was limited to 1 kDa. A database of all possible, but plausible chemical compositions, was created using the HR2 software tool by Kind *et al.*[320]. It was set to generate all compositions up to 1kDa, using the chemical rules (LEWIS and SENIOR) and probabilities (element/mass and element/element ratios) outlined in their publication. The maximum and minimum atom-counts were: C(1-100), H(1-150), N(0-50), O (0-50), P(0-9) and S(0-14).

The correct function of all algorithms was always checked throughout development. All critical computations were also repeated independently and compared to the algorithm's output.

IV.3. Results & Discussion

This chapter has two aims: to provide a method to identify molecules of biological origin, and to annotate them. By growing an organism on isotope enriched growth medium, all synthesized metabolites will be isotopically labelled, and so distinguished from chemical background. Furthermore, the increase in molecular weight as a result of isotope enrichment is proportional to the number of labelled atoms. For example, glucose ($C_6H_{12}O_6$) will shift from 180 Da to 186 Da. This is very valuable for annotating an empirical formula to unidentified peaks, as the number of atoms can be used as a search constraint in *ab initio* calculations to reduce the number of possible results.

The results and discussion is divided into three major sections: 1. *In silico* considerations regarding the usefulness and limitations of atom-count constrained searches and *ab initio* formula calculation. 2. The development of an automated workflow to analyse isotope labelled metabolite mixtures, acquired by LC-MS and using the gained atom-count information for molecular formula generation. 3. Testing this workflow on an *E. coli* dataset.

IV.3.1. *In silico* considerations for enhanced molecular formula calculations by isotope label analysis

Before any experiments were conducted, exhaustive *in silico* analyses were performed, to give a rigorous assessment of the value of atom constraints for *ab initio* molecular formula calculations. The focus is on elements that have stable isotopes available and so could readily be enriched in real systems (CHNOS). They should also be introducible into *in-vivo* systems, without compromising physiological function (CNOS). Furthermore, ideally they should be affordable in the required quantities (CNS).

IV.3.1.1. Knowing the carbon, nitrogen and sulphur atom count allows unambiguous annotation at ToF mass accuracy

A database query or elemental composition calculation of an unknown is usually based on its measured mass. The more accurate, the higher the chance to obtain an unambiguous hit. Mass accuracy is expressed as a relative measure between the difference of measured and calculated mass, according to equation IV.1. For convenience, it is expressed as parts per million (ppm).

$$Error_{ppm} = \frac{(mass_{measured} - mass_{calculated})}{mass_{calculated}} \times 10^6 \quad [ppm] \quad (IV.1)$$

The following simulation aims to find out how accurate a mass spectrometer would need to be, to prevent such ambiguities, *i.e.* being able to differentiate a composition from its most similar weighing neighbour. Furthermore, to what degree knowing the exact atom count of a particular element can afford greater uncertainty of the mass measurement.

Three major databases were used: (1) The manually curated *E. coli* pathway database EcoCyc [312, 313] ($n = 1,188$ compounds), (2) the National Institute of Health's (NIH) PubChem database [56] ($n = 212,910$) and finally (3) a database containing all plausible molecular formulas up to 1 kDa, generated *in silico* by a freely available HR2 software tool from Kind *et al.*[320] ($n = 30,914,124$). This gives three different scenarios:

(1) If a mass is matched against EcoCyc, this resembles a "naive" database search. Only molecules already known for an organism are expected as hits. As this is the scenario with

the least number of possibilities out of the three, the instrument accuracy requirement to deliver unambiguous hits should be the lowest one.

(2) A mass is matched against PubChem. In this scenario, the required accuracy is calculated to unambiguously assign a compound against all compounds ever reported.

(3) The ultimate test. Unambiguously assigning a compound against the background of all possible elemental compositions. This scenario is effectively equivalent to an *ab initio* calculation of the molecular formula.

It is important to point out that all molecules from (1) are contained in (2), and (2) is a subset of (3) (with a few rare exceptions).

In order to test these different scenarios, it was decided not to choose specific compounds as models [234], but rather to perform larger-scale simulations, to deliver more generally applicable conclusions. Therefore, all molecular formulae contained in Ecocyc were queried against (I) itself, (II) PubChem and (III) all possible elemental compositions. The required level of instrument accuracy ($\log_{10}$ ppm) for unambiguous assignment of the remaining fraction of all queried compositions is shown in figure IV.2a for scenario (I), figure IV.2b for (II) and figure IV.2c for scenario (III).

The slope may be interpreted as a cumulative probability density function for the likelihood to resolve a compound at a given ppm accuracy, and therefore changes from scenario to scenario. As expected, the mass accuracy required to resolve the same fraction of queried compounds rises with the increasing number of possibilities ((I)→(II)→(III)). Even without additional atom constraints, all compositions could be differentiated with a state-of-the-art ToF detector (3 – 5 ppm) in scenario (I). This is still true for 80 % of the compounds in scenario (II), only around half of the formulae could be calculated *ab initio* at 1 ppm (III). Clearly, constraining the atom-count during the search improves the probability of obtaining unambiguous hits. Unfortunately, the resolution of the graphs is rather coarse. A higher resolution would allow much clearer assessment of the impact of atom-count constraints.

As the graph resolution is based on matching pairs between two datasets, it can be im-

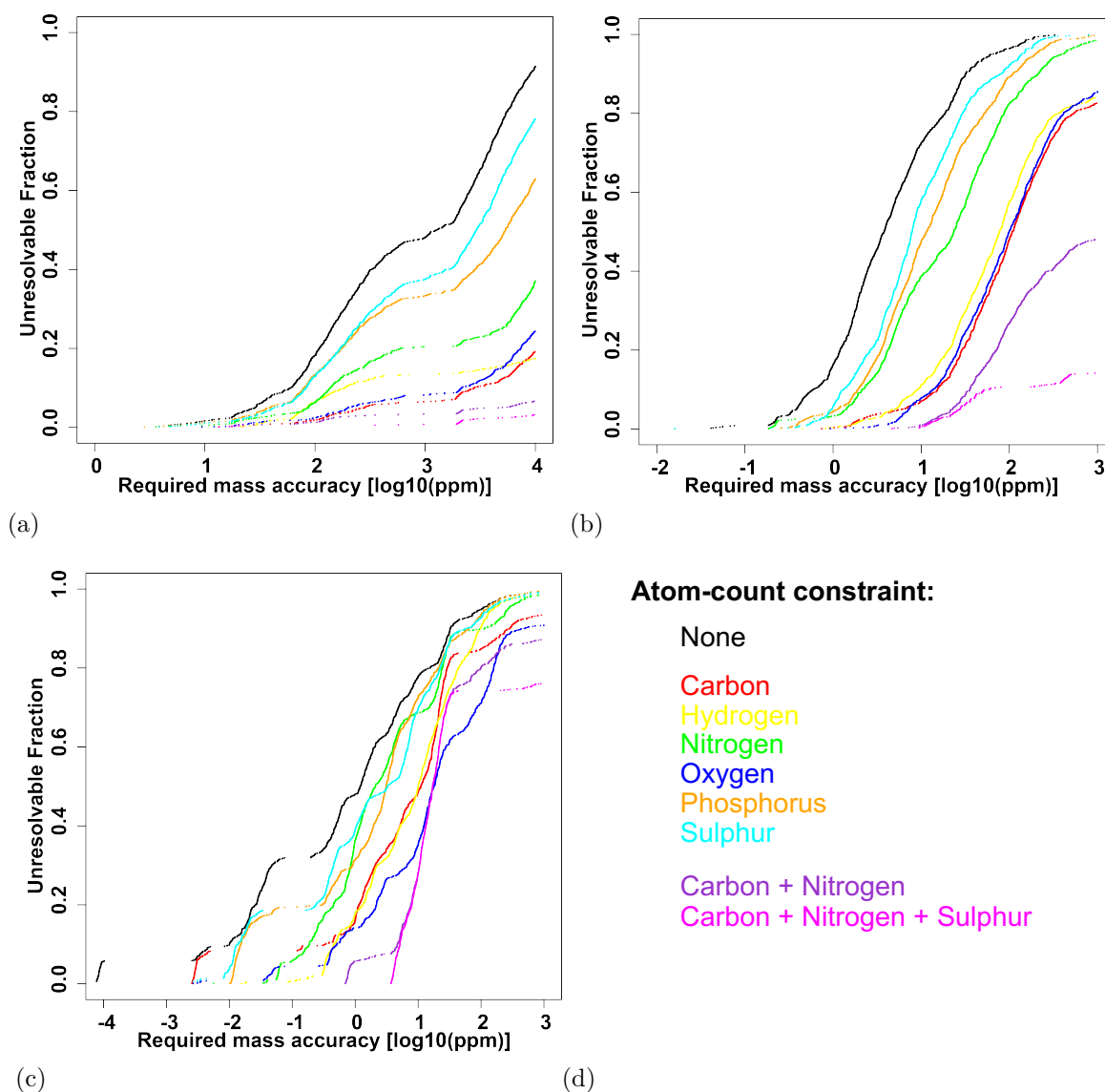


Figure IV.2.: **Instrument accuracy requirements impacted by various atom count constraints I:**

The fraction of unresolved unique elemental compositions is plotted against the required instrument accuracy (in ppm) to resolve those. The ppm domain is on a logarithmic (base 10) scale. In figure (a - c), the EcoCyc dataset is matched against (a) itself, (b) PubChem and (c) all possible combinations.

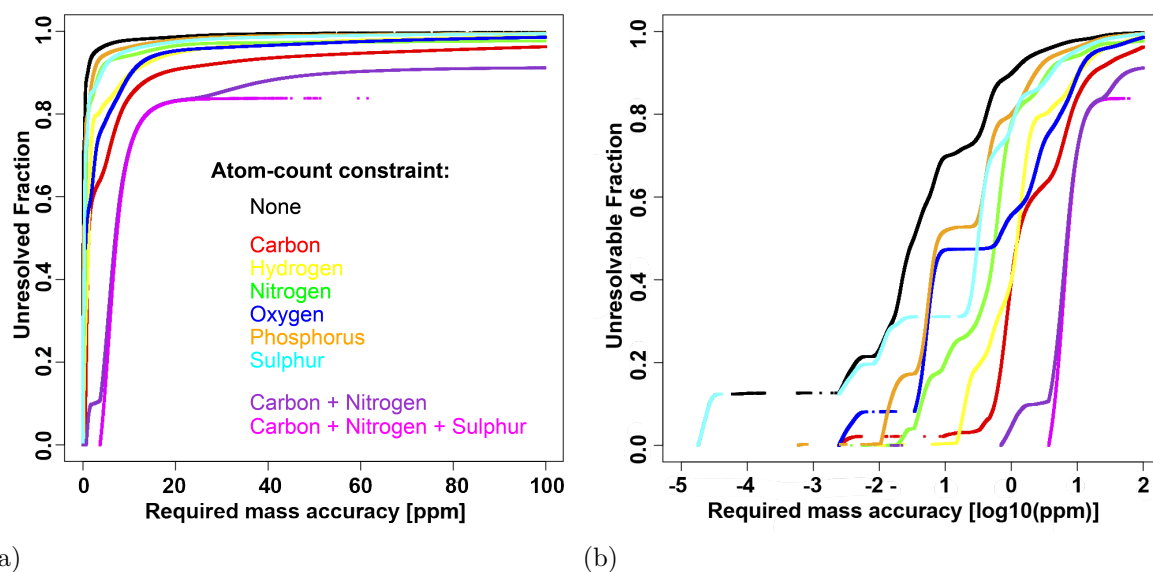


Figure IV.3.: **Instrument accuracy requirements impacted by various atom count constraints II:**

The fraction of yet (un)resolved unique elemental compositions is plotted against the required Instrument accuracy (in ppm).

In figure (a) the ppm domain is on a native, in figure (b) on a logarithmic (base 10) scale. In both cases, elemental compositions from the PubChem dataset are matched against all possible compositions. For more information, please refer to the text.

proved by increasing the number of matching events. As the main goal is to perform *ab initio* molecular formula generation it was decided to focus on scenario (III). However, rather than just matching the small EcoCyc dataset onto all possible molecular compositions, it was decided to use all chemical compositions contained in the PubChem database. The results can be seen in figure IV.3a and IV.3b without and with a logarithmic ppm scale, respectively. The graph with the native ppm scale displays the power of high mass accuracy instrumentation. Above about 10 ppm resolution, it is effectively not possible to unambiguously calculate molecular formulae. In order to determine a majority of all compositions unambiguously, and without using any atom-count constraints, one would require instruments at least two orders of magnitude more accurate than the current state-of-the-art. In order to resolve all compositions, the requirement would even be extended by 5 orders of magnitude to one part per quadrillion (ppq) accuracy. Kind *et al.* in 2006 [319] and Green *et al.*[209] also concluded that, instead of focusing on improving mass accuracy beyond one ppm, it is more valuable to use other information, such as isotope

patterns (although they did not provide the same exhaustive comparison to all possible backgrounds).

Atom-counts provide additional information. Indeed, as figure IV.3b shows, certain constraints are very powerful when combined with high mass accuracy measurements, with the most and least powerful being carbon/hydrogen and sulphur, respectively. This is as expected as carbon/hydrogen display the highest diversity in the number of atom-counts, whereas sulphur is almost always present as 0 - 2 atoms per compound. Both hydrogen and carbon constraints could resolve all compositions at around 0.1 ppm accuracy, whereas sulphur has little benefit over no constraint at all. Knowing the atom-count of multiple elements is more powerful than for single elements. Carbon/nitrogen (CN) and carbon/nitrogen/sulphur (CNS) are of particular interest, as these isotopes can readily be introduced into a biological system, at reasonable cost. The simulations show that CN and CNS can unambiguously resolve all compounds below 1 kDa at 0.015 ppm and 3.7 ppm, respectively. This means that knowing the number of carbon, nitrogen and sulphur atoms, plus molecular mass measured with a current state-of-the-art time-of-flight mass spectrometer, it is possible to resolve all possible molecular compositions.

Graphs for all double and selected higher constraints can be found in figure C3.1 in the appendix.

Bowen *et al.*[58] compared the use of atom-count constraints to the use of isotope pattern analysis for *ab initio* molecular formula generation. They focused on the carbon and nitrogen double constraint at 5 ppm mass accuracy. They also compared the "all compounds formula database" generated by the HR2 software tool, used in this thesis, with a brute force formula generator that generates every combinatorial molecular composition, even if chemically implausible. They found that at 5 ppm and carbon/nitrogen atom-count constraint, 55 % of PubChem's compositions could be unambiguously calculated when matched against the HR2 generated list. This dropped to 50 % for the all combinatorial possible formulae. In this thesis's simulation the same scenario yields 75 % unique solutions. Even though this difference sounds large, it is in fact not that different: the CN curve (in figure IV.3b) has its steepest slope between 10 and 4 ppm. It falls from 29 % unique compositions at 10 ppm, to 87 % within just 6 ppm. Thus, small differences in simulation parameters, such as their inclusion of compositions up to 1244 Da, could easily

cause these differences. Hence, the analysis here is consistent with previous results. It also shows that pure combinatorial molecular formula generation, *i.e.* without any probabilities or chemical rules, does not result in vastly different results from the approach taken here. Overall, Bowen *et al.*[58] found atom-count constraints to be superior to isotope pattern analysis for *ab initio* molecular formula generation, particularly for compounds above 500 Da.

IV.3.1.2. Impact of molecular weight for unambiguous molecular formula calculation (considering different atom-count constraints)

One would also expect a molecule's mass to be an important factor when examining the required instrument accuracy to calculate its molecular formula and the impact of atom-count constraints - the lower the molecular weight of a compound, the more likely a database query or *ab initio* formula calculation will deliver unambiguous hits. Indeed, based on their experiments, Rodgers *et al.*[501] hypothesised that carbon-count constraints could triple the upper mass limit for unambiguous assignment. Hegeman *et al.*[234] use poly(phosphoserine) ($n = 1, 2, 3, \dots, 10, 15$) as a model for compounds that are difficult to resolve, due to their mass being close to the mode (top) of the mass distribution around an integer mass bin. Using this model, they predicted that at 3 ppm instrument accuracy, molecules can be unambiguously resolved up to a molecular weight of 186 if unconstrained, 278 with known carbon-count, 214 with nitrogen constraint and 422 if the number of carbon and nitrogen atoms is known.

Therefore this thesis tries to explore the general dependency between molecular weight and unambiguous assignments, considering different atom-count constraints. Using the *ab initio* calculation scenario from above, all compounds in PubChem were queried against all possible molecular formulae, generated by a software tool from Kind *et al.*[320]. However, to reduce computational requirements, only compositions that cannot be resolved at 10 ppm or less are considered, in contrast to the 100 ppm used in the simulations above. At 10 ppm accuracy, there still is no unique molecular formula for 97.9 % of all masses contained in the PubChem dataset.

Figure IV.4 shows individual data points for each molecular composition, rather than ranking them with ascending ppm accuracy requirement, as done in the section above. The results agree with the mass limits for unambiguous assignments (at 3 ppm) proposed by

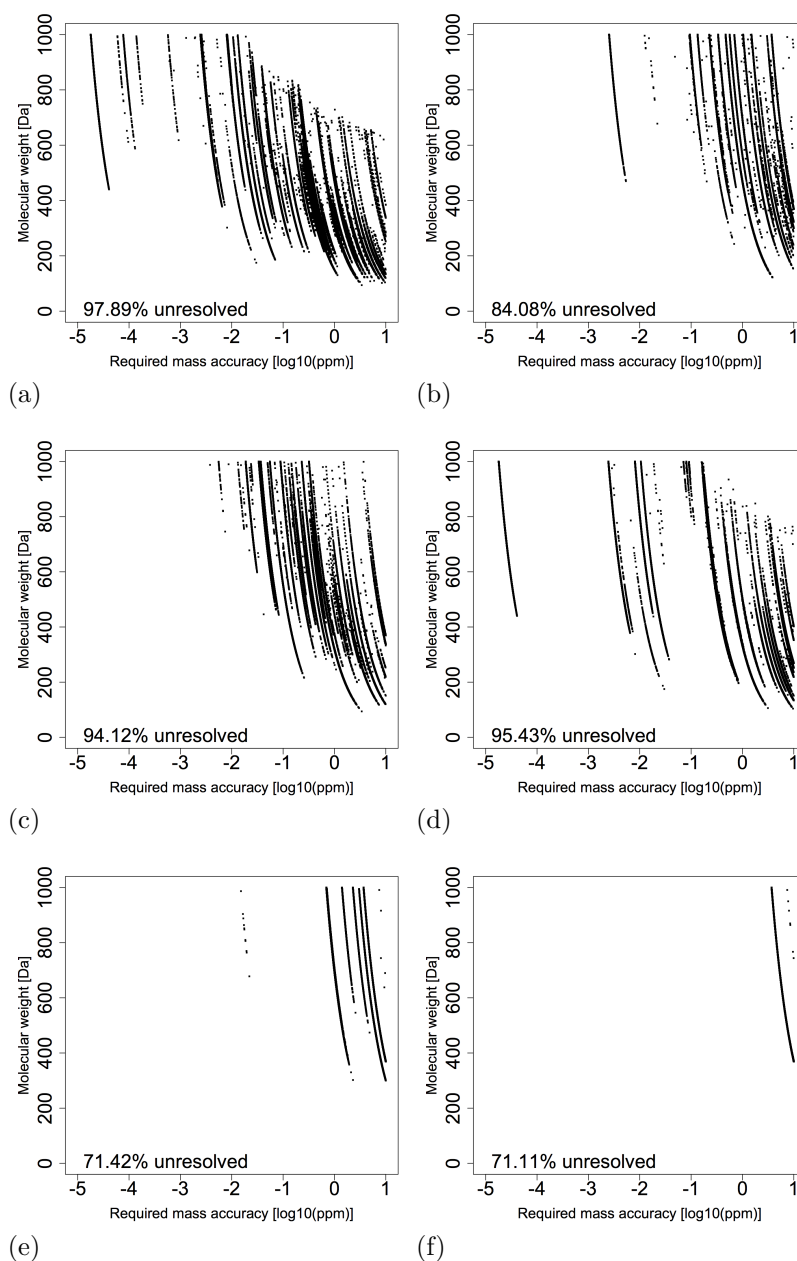


Figure IV.4.: **Required accuracy to determine molecular composition versus its mass :**

Each datapoint represents a molecular composition in the pubchem dataset. The (log 10 ppm) instrument accuracy required to be able to unambiguously calculate its composition from its measured mass is plotted against its mass.

Figure IV.4a: No atom-count constraint

Figure IV.4b: Carbon-count constrain

Figure IV.4c: Nitrogen-count constraint

Figure IV.4d: Sulphur-count constraint

Figure IV.4e: carbon- and nitrogen-count constraint

Figure IV.4f: Carbon- nitrogen-, and sulphur-count constraint.

The percentage below shows the percentage of all data points in the plot from the whole PubChem dataset.

Hegeman *et al.*. However, the ability to unambiguously resolve a compound's composition does not depend on its molecular weight as much as might be expected. Indeed, without atom-count constraints and at ToF accuracy, many compositions with less than 200 Da molecular weight may not be resolved, but many above 600 Da will be.

An investigation into what actually determines the required accuracy was conducted. It was found that it depends on the atom-count ratio of the two matched compositions (for any given element). There are certain ratios, which are much harder to resolve than others. A more detailed investigation would exceed the scope of its thesis. It can be concluded, however, that the required accuracy does not depend on the absolute mass *per se*, but also on the exact stoichiometry of the matched molecular compositions.

IV.3.2. A protocol for stable isotope labeling of carbon, nitrogen and sulphur in *C. elegans*

Introducing atom constraints for C, N, and S means that all elemental compositions up to at least 1 kDa can be generated *ab initio*. Therefore a protocol to introduce those elements into *C. elegans* was developed.

In order to isotopically label *C. elegans*, its food source, *E. coli*, must be labelled. Two protocols for stable isotope enrichment in *C. elegans* exist. One for ^{13}C by An *et al.*[15] and one for ^{15}N by Krijgsveld *et al.*[337]. However, neither was suitable for use here (and the second was not published at the time of this experiment). An *et al.* use the OP50 *E. coli* strain, which is a uracil auxotroph. Full isotope enrichment would therefore require custom synthesis of $^{13}\text{C}/^{15}\text{N}$ labelled uracil, which was not feasible. Krijgsveld *et al.* use commercial ^{15}N growth medium, which is not available for all isotope combinations in this study. Therefore a new protocol was developed.

The metabolically ÔcleanÕ *E. coli* wild type strain NCM3722 was chosen, which is widely used in metabolomics studies [29, 375]. NCM3722 grows on completely synthetic minimal medium, for example the one developed by Gutnick *et al.* [218]. However, adaptation of this protocol has previously been reported for ^{13}C and ^{15}N stable isotope enrichment, but not for sulphur [375]. To this end all salts with sulphur-containing counter ions, such as sulphates, were replaced by their chloride equivalents.

Isotope enriched sulphur can be conveniently introduced through addition of Na_2SO_4 , with $\text{Na}_2^{33}\text{SO}_4$ being commercially available. Due to its high price, NCM3722's minimum sulphur requirement in the used medium was determined by OD₆₀₀ measured growth curve experiments with different Na_2SO_4 concentrations. Results, displayed in figure IV.5 indicate, that a sulphur concentration of 10 mg per litre of medium is sufficient, to not constrain bacterial growth. This would correspond to 45 mg per litre $\text{Na}_2^{33}\text{SO}_4$. Indeed, this result is similar to the one determined by Cowie *et al.*[113] for growth of a not further characterised *E. coli* strain in M9 minimal medium (equivalent to 5 mg sulphur per litre).

The protocol generates carbon, nitrogen and sulphur labelled *E. coli*. To feed *C. elegans*, these are pelleted and streaked onto agar plates, which are prepared without the use of (unlabelled) peptone. Apart from *E. coli* as food source, *C. elegans* requires the addition

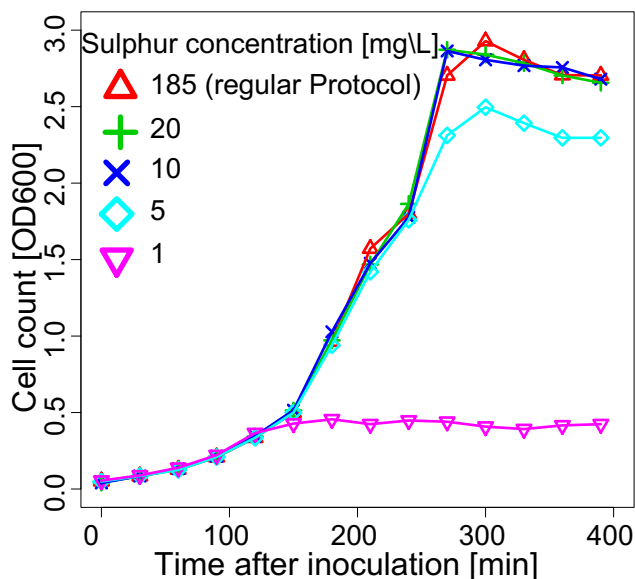


Figure IV.5.: **Minimum sulphur requirement for *E. coli* strain NCM3722:**

The minimum sulphur requirement for *E. coli* strain NCM3722 in Gutnick minimal medium is determined. 1 mg/L shows a serious constraint, 5 mg/L a smaller one. 10 mg/L supports normal growth.

of cholesterol to the growth medium, as *C. elegans* lacks enzymes required for the biosynthesis of sterols from isoprene [339]. Fully labelled ^{13}C cholesterol is not commercially available. As *C. elegans* lacks the capability to synthesise sterols, it is also not capable of breaking down supplemented ^{12}C cholesterol to "contaminate" the growth medium. Thus, natural abundance cholesterol may be used. The cholesterol is introduced by solubilisation in ethanol and addition to the agar, prior to pouring plates. Even though ethanol was allowed to evaporate from the poured Petri dishes by leaving their lids open until solidification [580], residual ethanol remained in the medium. Metabolism of the remaining ethanol by the bacteria or worms led to an incomplete labelling of *C. elegans* biomolecules. Unfortunately, this rendered the first *C. elegans* dataset, which was intended for an in-depth *C. elegans* metabolite characterisation, unusable for automated isotope assisted molecule annotation. Figure IV.6a shows an example mass spectrum of an incompletely labelled lipid. To avoid this problem, cholesterol was added to the agar directly (rather than as a solution), with dispersion of the cholesterol as microcrystals by shaking and extensive sonication. Worms were grown over several generations on ethanol-free growth medium to ensure there were no cholesterol depletion defects, such as abnormal development or premature death [413, 151]. Worms cultured in this manner, did show a complete ^{13}C la-

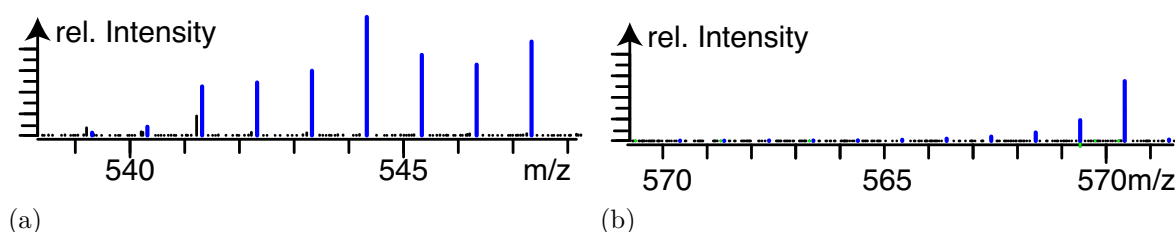


Figure IV.6.: **Isotopic abundance pattern of (in)completely labelled lipid:**

Both figures show the ^{13}C mass spectrum of a lipid. In figure, IV.6a, the labelling is incomplete, leading to a mixture of species with various levels of isotope incorporation. The labelling in figure IV.6b, however, is considered complete. It could be described as an inversion of the usual ^{12}C abundance pattern (1 % ^{13}C). This is due to the ^{13}C glucose used for enrichment was obtained at 99 % isotope purity and therefore has 1 % ^{12}C content.

bellung of their biomolecules (considering the purity of commercially available ^{13}C glucose), as shown in the mass spectrum in figure IV.6b.

IV.3.3. Development of a software algorithm for automated data reduction and molecular formula generation by stable isotopes

This section describes development of a fully automated software-engine that can extract atom-count information from LC-MS data of a biological system that has been enriched with stable C, N and S isotopes. It uses the extracted atom-count information to generate molecular formulae, to accelerate assignment and help differentiate chemical contamination from real biological information.

Preliminary experiments with just carbon and nitrogen labelling showed an abundance of non-biological contaminants. Indeed, those signals, which are shared between different isotope labelling regimes, were enough for xcms to perform retention time alignment. Thus it was decided not to additionally spike samples with unlabelled extract.

E. coli data were used for developing the algorithm. Bacteria were grown according to the protocol developed in section IV.3.2. Thus these datasets also provided a test for the carbon, nitrogen, sulphur labelling protocol. *C. elegans* labelling was attempted but the samples had to be discarded because of some metabolites being ^{12}C "contaminated", due

to the use of ethanol. Those samples can still serve as a reference for annotation, but method development was constrained to the *E. coli* samples.

As shown in table IV.6, the samples were run in four different conditions: reversed phase and hydrophobic liquid interaction (HILIC), each in positive and negative ionisation mode, to assure the developed workflow will perform with a range of chromatographic and electrospray conditions.

	HILIC	Reversed Phase
ESI+	HILIC+	RP+
ESI-	HILIC-	RP-

Table IV.6.: **Dataset for algorithm development and validation**

The algorithm can be used with the output from any LC-MS processing software, but was developed here on retention time aligned (retention time $\times$ mass) peak lists generated by xcms. The algorithm requires one natural abundance sample, plus samples under 5 different stable isotope labelling conditions: singly labelled with carbon (C), nitrogen (N) and sulphur (S), doubly labelled with nitrogen and carbon (CN), and triply labelled with carbon, nitrogen and sulphur (CNS). The algorithm makes no prior assumptions about the data, except an estimation of the retention time deviation (rtE) and mass errors (ppmE). As will be shown in section IV.3.4.1, those may be optimised for best possible results. The workflow of the algorithm itself is outlined in figure IV.9.

In the first step, peaks are queried whether they are biological in origin, and if so, their number of carbon, nitrogen and sulphur atoms is determined (figure IV.7b). To this end, a peak in the unlabelled sample must be matched to the corresponding peaks in the labelled samples (figure IV.7a). This matching process equates to the grouping step of retention time alignment. However, it is slightly more complicated, as a labelled peak will have a different mass. According to equation IV.2, the mass difference (Δm) between the unlabelled and labelled mass, equals to the labelled element's atom-count (n) times the difference (Δu) in atomic weight between the main natural abundance atom and its stable isotopologue of the element in question.

$$m_{native} - m_{isotope} = \Delta m = n \times (u_{native} - u_{isotope}) = n \times \Delta u \quad (\text{IV.2})$$

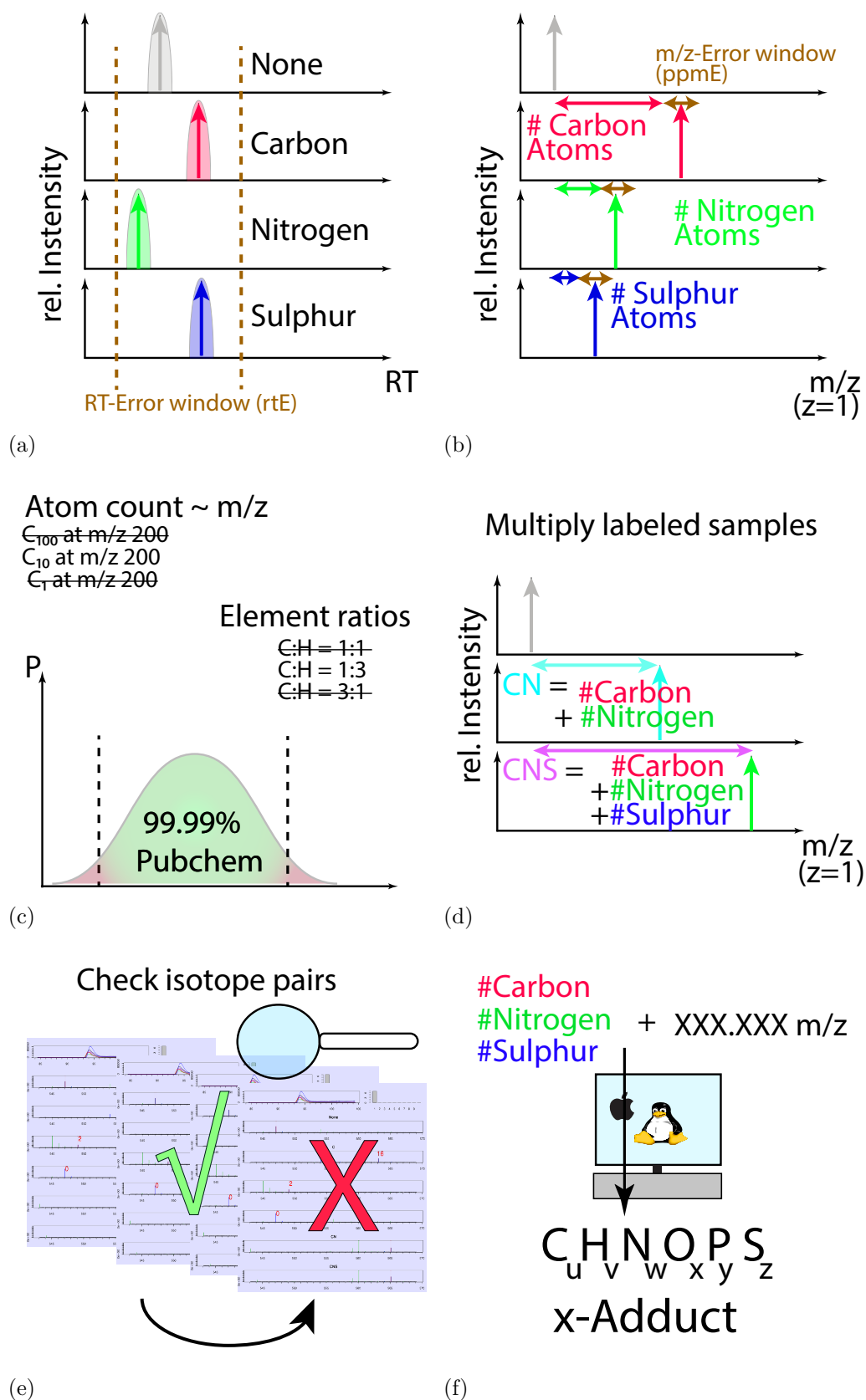


Figure IV.7.: **Workflow of data reduction and annotation algorithm:**
Details are explained in the text.

Thus, the algorithm will flag every pair (inside a retention time window) that is $n \times \Delta u$ Dalton apart, as a potential match. Δu is atom specific as shown in table IV.7. This is because all atoms have a specific mass defect, *i.e.* the mass converted into binding energy is different for each atom. Thus, the mass difference Δu between ^{12}C and ^{13}C , for example, is not simply a unit mass, but a very specific value, depending on the mass defects of ^{12}C and ^{13}C .

Element	Mass [Da]	Isotope Mass[Da]	Difference Δu [Da]
Carbon	12.000 000	13.003 354	1.003 355
Nitrogen	14.003 074	15.000 108	0.997 035
Sulphur	31.972 071	32.971 458	0.999 388

Table IV.7.: **Exact mass and difference (considering mass defect) of used isotopes [132, 624]**

A (reflectron) ToF mass spectrometer’s accuracy is high enough to discern these mass differences. Thus, the algorithm can distinguish whether a shift is due to the presence of an isotope or a coeluting molecule of higher mass value. To make the most of this effect, the algorithm requires an estimate of the mass accuracy of the acquisition, measured in ppm (ppmE). After this step, a list of unlabelled - labelled peak pairs is outputted.

The mass defect helps to reduce possible peak pairs. Still, due to the sheer number of combinatorial possibilities, particularly considering the presence of natural abundance isotopes in unlabelled samples (*e.g.* ^{13}C abundance in a ^{12}C system) or the presence of unlabelled molecules in stable isotope enriched samples (*e.g.* ^{12}C "contamination" in a ^{13}C system), the number of proposed pairs is very high. The list could be drastically truncated by querying the probability of those pairs. To this end, the PubChem database was used to define confidence intervals on the basic architecture of molecules, similarly as proposed by Kind *et al.*’s empirical "7 Golden Rule" analysis [320]. As shown in figure IV.7c the molecular weight to atom-count ratio (with carbon and hydrogen in particular) and elemental ratios were used. In all cases, a 99.99 % confidence interval was used. Thus, a molecule would need to have a molecular weight to atom-count ratio at least as extreme as 0.01 % of all compounds contained in PubChem, to be discarded. Even though this algorithm’s goal is to annotate new, non-database metabolites, it is highly unlikely for a biological metabolite to exceed these boundaries - particularly as PubChem also contains synthetic molecules,

such as C_{60} . For example, to pass this rule, a molecule with a mass of 100 must contain from one to 8 carbon atoms ($8 \times 12 = 96!$). Therefore this step just leaves pairs that may lead to biologically plausible molecules. Atom-count to molecular weight and atom-count to atom-count ratios can be found in the appendix, in table C3.1 and C3.2, respectively.

In the next step, the algorithm combines these pairs by comparing them to multiply labelled samples, as shown in figure IV.7c. A carbon and a nitrogen pair, which both share the same unlabelled counterpart, are required to have a matching carbon-nitrogen double label in their retention time window. For example, in the case of $C = 7$ carbon atoms and $N = 1$ nitrogen atom, the carbon-nitrogen peak would be required to be $CN = C + N = 7 + 1 = 8$ Da higher than the unlabelled sample. Again, exact masses, with specific mass defects are used. The carbon-nitrogen double pairs together with the sulphur pairs are then matched against the carbon, nitrogen and sulphur triple labels ($CNS = C + N + S$). Apart from Hegeman *et al.*[234], none of the other published methods use such multiple constraints, even though they are powerful at reducing combinatorial possibilities - as will be shown in the next section.

At this stage, it would already be possible to calculate the molecular formula of the peaks giving rise to the *None/C/N/S/CN/CNS* pairs. However, the aim of the workflow is to rigorously and accurately reduce the chemical information to biologically relevant, non-redundant data. Thus, the operator is given the chance to check the validity of the correlated pairs. To this end, check tables for each pair are generated in the PDF file format. A check table, such as displayed in IV.8 has three major elements. It contains the overlaid extracted ion chromatograms (EIC) of the pairs. It also displays the individual mass spectra for each labelling regime, extracted at the apex of each EIC. Inside the mass spectra, the peaks of the pairs are highlighted and the proposed atom-count displayed. Also, the potential isotope peaks are coloured differently from the rest of the mass spectrum's peaks. Finally, a charge indicator indicates the charge of a (potentially) multiply charged pair. Even though the formula generator considers multiply charged ions, there as yet is no support to deconvolve multiple charge envelopes [665], as those play a subordinate role in metabolite profiling.

The check tables were found to be extremely practically useful for evaluating the experimental data. They allow one to spot and discard mis-matched pairs as well as those which

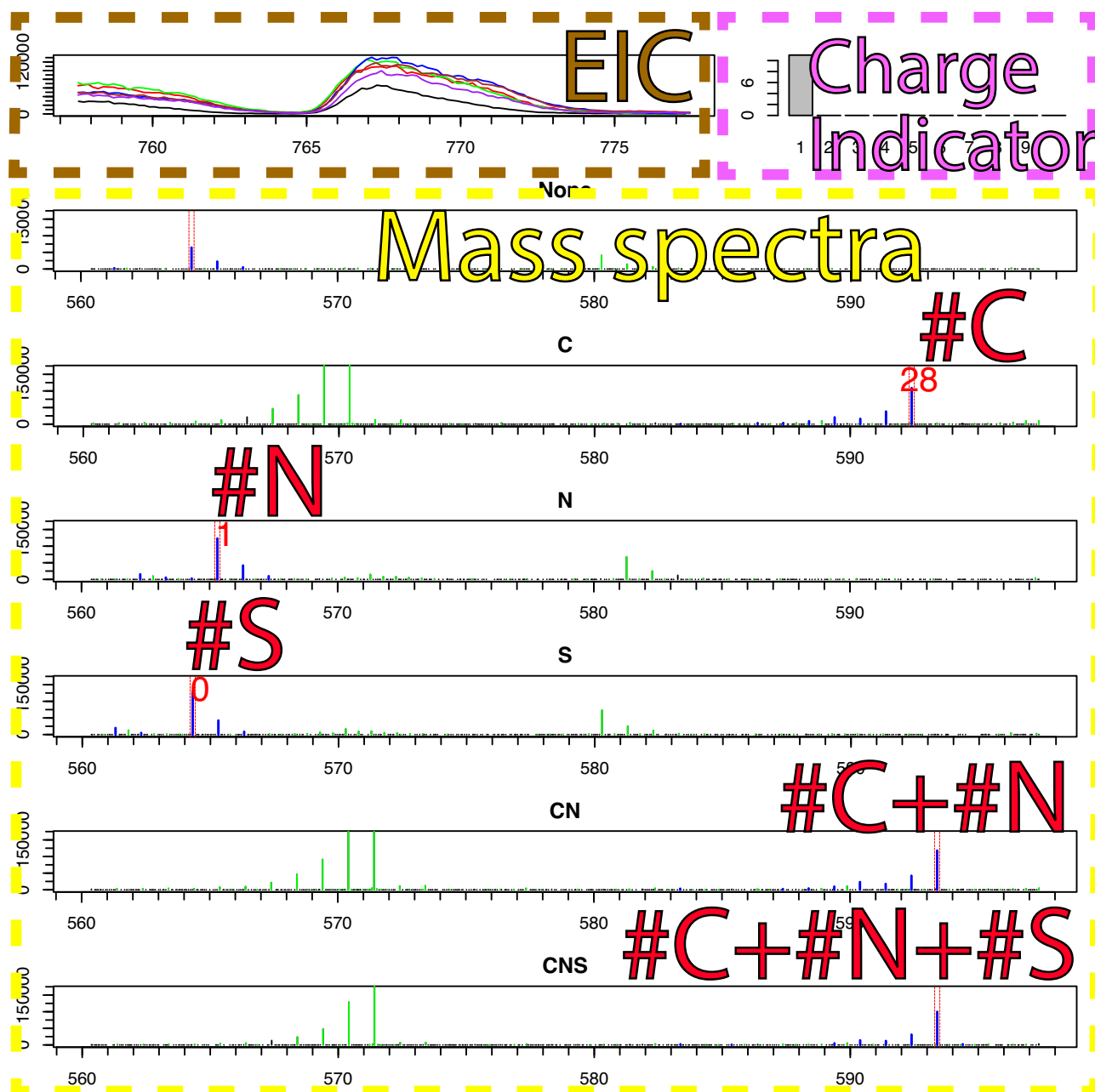


Figure IV.8.: Check tables for isotope peak pairs:

The check table, consists of three elements: The overlaid EICs, the individual mass spectra at EIC apex and a charge indicator. In this a case a pair with 28 carbon, one nitrogen and zero sulphur atoms is proposed. As the EICS are aligned perfectly, no multiple charge is present and the correct peaks were detected inside the mass spectra, this assignment looks plausible (compare to figure IV.9b). Another C, CN and CNS labelled compound seems to be co-eluting at 570 Da. However, it is coloured green. All isotopes which according to accurate mass measurements belong to the proposed isotope pair, are coloured blue. This means the signal at 570 does not belong to the proposed isotope pairs. Signals not identified as peaks by the xcms peak detection step are coloured black.

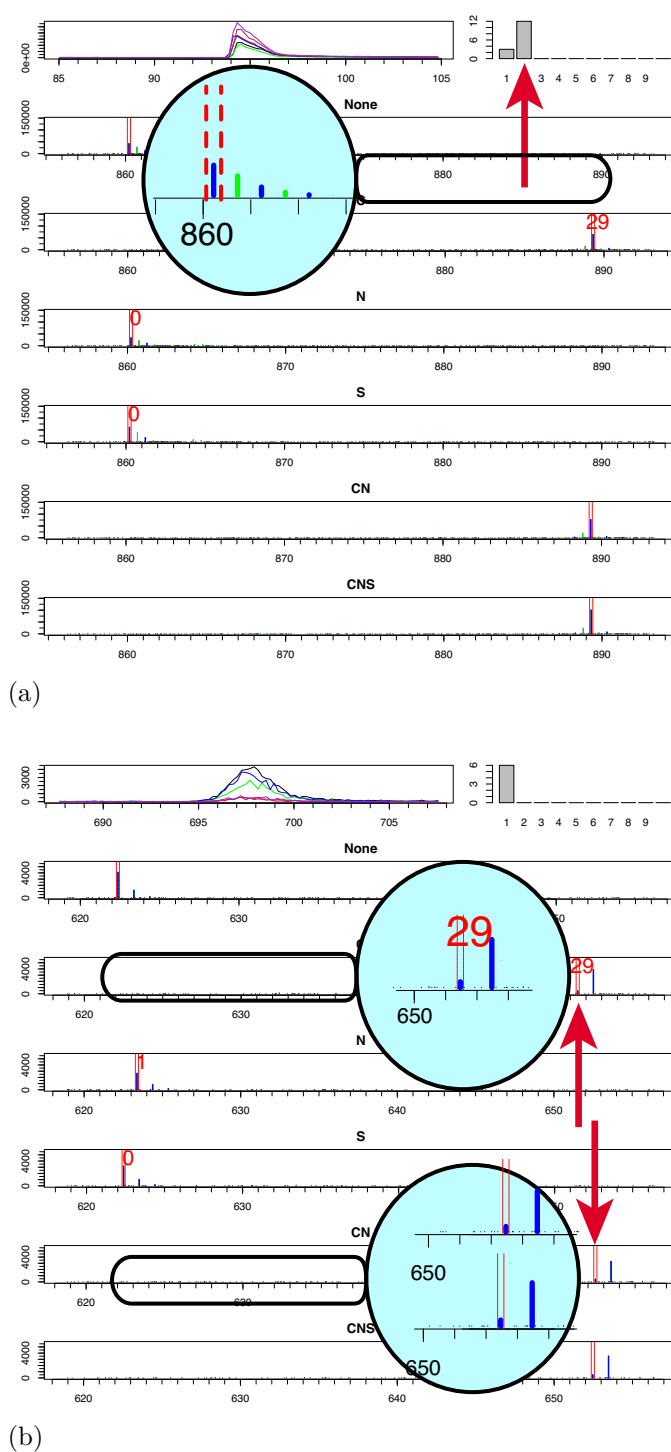


Figure IV.9.: **Multiply charged and isotope pair mismatch:**

Figure IV.9a shows an isotope pair, which is doubly charged. This is detected by the charge indicator in the top right corner. Indeed, as shown in magnification the peaks have a distance of around 0.5 Da, rather than 1 Da. Figure IV.9b shows a mismatched isotope pair. For the C, CN, CNS the M-1 isotope was used, leading to an underestimation of the atom-count.

do not meet the user's quality criteria, such as incomplete chromatographic separation of isomers. Figure IV.9a and IV.9b show an example of multiply charged isotope pairs and a mismatch, respectively. Mismatches in which the M+1 or M-1 isotope was selected, were found to be the most common mismatch by far. Thus it was started to fit isotope abundance patterns [259], to recognise and remove such events. However, this is non-trivial as often different metabolites display a different degree of isotope label incorporation. Therefore this algorithm's rules had to be extended to a level that it also flagged oddly labelled, but correct matches. To guarantee comprehensive and high quality assignment, all algorithm predictions in this thesis were manually checked that they were not mis-matches.

After manual curation, a molecular formula is generated for each pair (figure IV.7f) under consideration of adducts. In previous studies, Giavalisco *et al.* [195, 196] did not consider any other adducts apart from $M^{+/-}$ and therefore are likely to have mismatched compounds during their database search. In their latest study, however, they considered three common adducts [194]. The algorithm here uses a comprehensive list of adducts [574, 44] during formula calculation: 30 and 14 adducts for positive and negative mode, respectively, as shown in table C3.3 in the appendix. All formula calculation is performed on the fly and no lookup databases are required. Many of the previous studies just constrained their searches to publicly available compound databases, not taking unknown compositions into account [196, 194]. The generation algorithm itself is an efficient re-written implementation of the open source HR2 formula generator [320]. This is the same formula generator which was used to generate the "all possible compositions" database for the simulations section.

In a final validation stage, it is checked whether the proposed molecular formula, including adducts, holds true for all labelling regimes.

Most of the code-handling is performed inside R. However, all computationally intensive functions, such as finding the isotope pairs, correlating them to multiple labels and the formula generation, were implemented in optimised C functions, so that the algorithm runs quickly. Exact timings depend on computer specifications, the number of detected peaks and the size of retention time and mass accuracy error windows, rtE and ppmE, respectively. However, in most cases the process was faster than the peak detection of a single sample, as performed by xcms in the first instance.

IV.3.4. Optimisation and testing of the data reduction and formula generation algorithm

This section aims to assess the algorithm's performance to reduce data and generate molecular formulae.

The algorithm consists of four subunits. Peak detection (1), data reduction (2), which is based on correlating peaks across runs (3), and the final formula generation (with consideration of multiple adducts) (4). For peak detection (1), the well-tested *xcms* [531] software package [566, 141, 62] can be used and therefore does not require further attention. The data reduction algorithm will be tested for its performance in how far it picks up all biological features, but does not discard them during the data reduction. Throughout the process (3), the algorithm aligns features from different labelling regimes, which allows for misalignments to occur. But as the user can check potential misalignments by inspecting the check panels produced by the algorithm, as described in the previous section, additional validation is not necessary. Correct function of the molecular formula generation with consideration of different adducts, will however, need to be tested.

The full *E. coli* dataset will be used for benchmarking and validation, to show that the algorithm works irrespective of chromatographic (RP/HILIC) and electrospray (+/-) modes.

IV.3.4.1. Testing and optimising the data reduction algorithm

Under ideal circumstances, the reduction algorithm strips the data of all non-biological noise and redundancy and retains all, and only the biological information.

The performance of any algorithm depends on the correct setting of its parameters. Thus, these should be optimized for best performance. During the design of the algorithm, care was taken not to make assumptions about the data and its parameters. However, two do need to be estimated; the error windows in the retention time (rtE) and *m/z* (ppmE) domain, which are used when matching the peaks of the different labelling regimes. If set too narrow, peaks that actually belong together may not be connected, leading to false negatives. If set too wide, unrelated peaks could be matched, increasing the number of false positives. Determining the best trade off between those two errors, for example by receiver operating characteristic (ROC) analysis [164], would require manual consideration

and classification of every possible peak-pair under the use of different thresholds. As this is not feasible, a similar, but more automatable approach was taken.

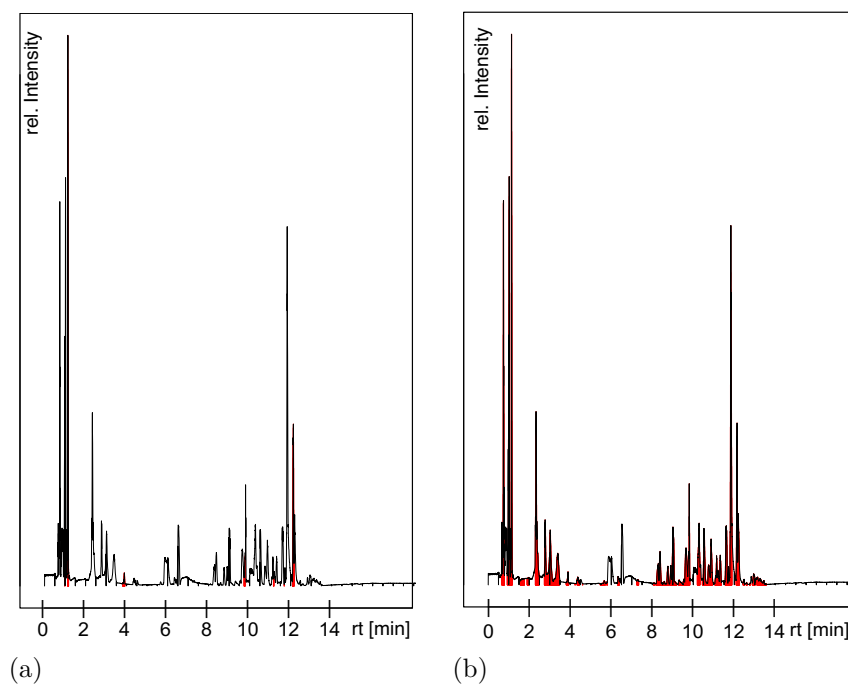


Figure IV.10.: **Optimising the retention time and m/z window during matching:**

Both figures display the base ion peak (BIP) chromatogram of the negative mode HILIC dataset. All peaks, which were identified to be of biological origin by the algorithm, are highlighted in red. Figure IV.10a uses a very narrow matching window ($rtE = 0.5s$, $ppmE = 1$), leading to very few assignments. The more liberal settings ($rtE = 10s$, $ppmE = 30$) in figure IV.10b lead to a very high number of "biological" signals. - Also false assignments when checked with the raw-data. Figure IV.11c shows the chromatogram of the ideal threshold for the sample set ($rtE = 2s$, $ppmE = 10$) in.

The rtE and $ppmE$ parameters were systematically varied between 0.25 and 10 s and 1 and 30 ppm, respectively, to observe the impact on the algorithm performance. The base ion peak (BIP) chromatogram's peaks were taken as performance indicators and as representatives for the whole sample set. This was done by visually highlighting the BIP peaks flagged as biological at a given $rtE/ppmE$ threshold. Figure IV.10a and IV.10b show the outcome of a low and high threshold, respectively. At low values, no peak was detected.

At very wide $rtE/ppmE$ settings, non-biological peaks - as confirmed by manual checks of the raw data - were included by the algorithm. Therefore, by manually comparing the differently thresholded BIP chromatograms with the raw data, the optimal settings for $rtE/ppmE$ could be selected. The corresponding chromatograms of the ideal settings for the different chromatographic and electrospray modes are displayed in figure IV.11.

Figure IV.11 shows a high amount of non-biological signal in the BIP chromatogram, which provides a challenge for the algorithm. It could be asked, whether the BIP generally is a good representation of all the peaks inside the sample set, and should therefore be used to estimate global parameters. The BIP displays the most intense m/z bin of each scan. However, as intensity information is not considered by the algorithm in any way, BIP signals should behave like lower intensity ones. Another analysis which considers all peaks was used to confirm that the visual BIP method yields good $rtE/ppmE$ thresholds. In figure IV.12 the number of total C-N-S pair combinations is plotted against the integral of the explained BIP area, at different thresholds.

In all four cases, the plot follows a similar pattern: an initial over proportional increase of BIP area versus peak pairs, followed by a reversal of this trend. Little gain in BIP area, but exponential increase in peak pairs. Even at the most liberal settings ($rtE = 10s$, $ppmE = 30$), the assigned BIP areas did not equal the full BIP area, despite the high number of C-N-S pairs. This indicates an explosion of false positive assignments at liberal parameter values. Also, when inspecting the plots, it becomes apparent that the ideal $rtE/ppmE$ thresholds should be conservative, to avoid false positives.

After determining the ideal rtE and $ppmE$ threshold for each dataset, the performance of the algorithm to reduce the number of signals, down to non-redundant biological information, was benchmarked. Table IV.8 shows the number of data points, at several stages of the algorithm.

The first thing to note is the overall 99.9 % reduction efficiency of the algorithm. Three stages of the algorithm show a particularly high performance: the use of probability based elemental ratios, the use of multiple labelling regimes and the semi-automated quality checks. The semi-automatic checks mostly eliminated biological features that were multiply charged. Thus, this need for manual curation does not indicate poor performance of

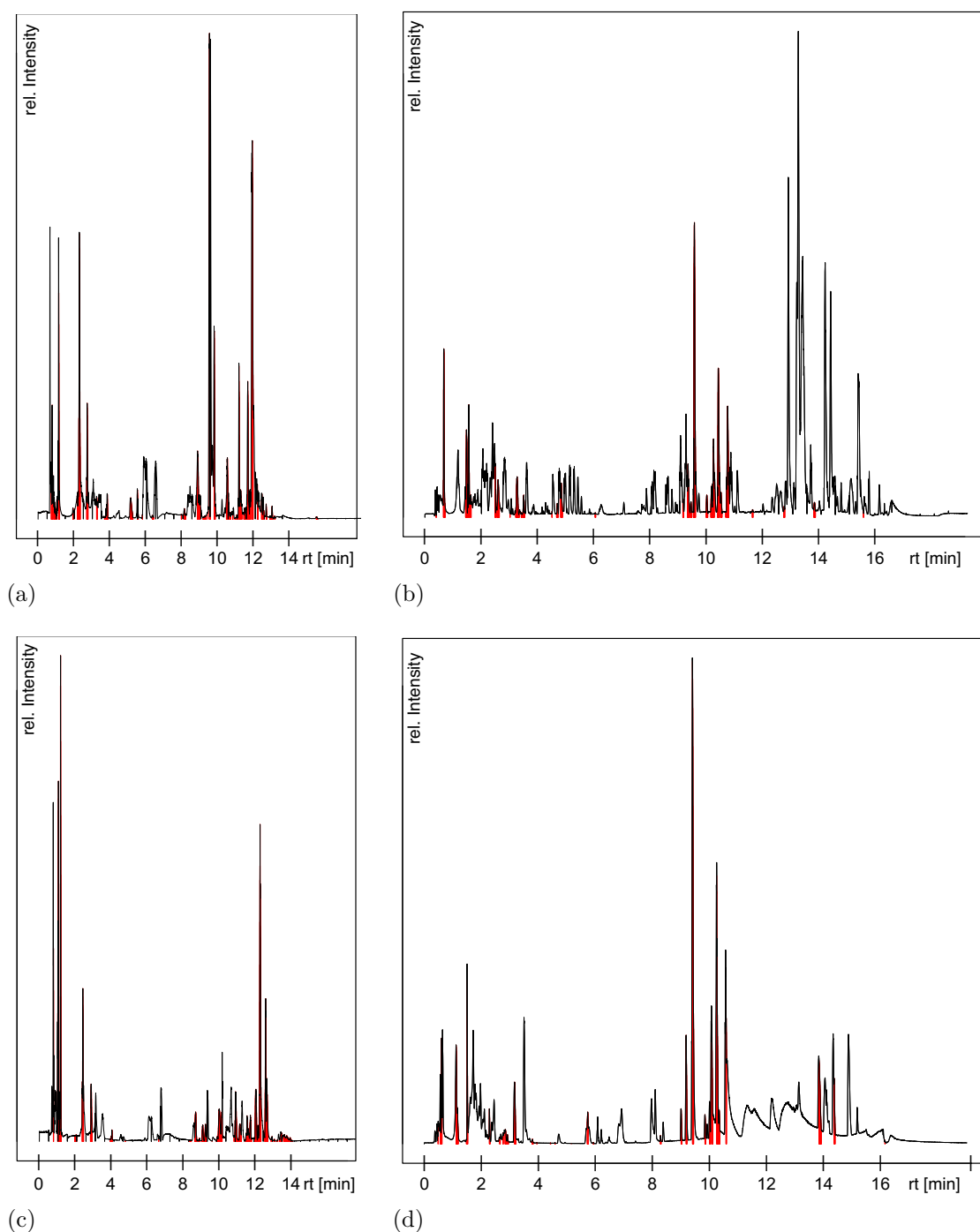


Figure IV.11.: **Ideal rtE and $ppmE$ settings for the four datasets:**

Figure IV.11a ESI+ HILIC ($rtE = 3s$, $ppmE = 6$).

Figure IV.11c ESI- HILIC ($rtE = 2s$, $ppmE = 10$).

Figure IV.11b ESI+ RP ($rtE = 1s$, $ppmE = 4$).

Figure IV.11d ESI- RP ($rtE = 1.5s$, $ppmE = 8$).

All peaks highlighted in red are of confirmed biological origin. The rest is unknown contamination.

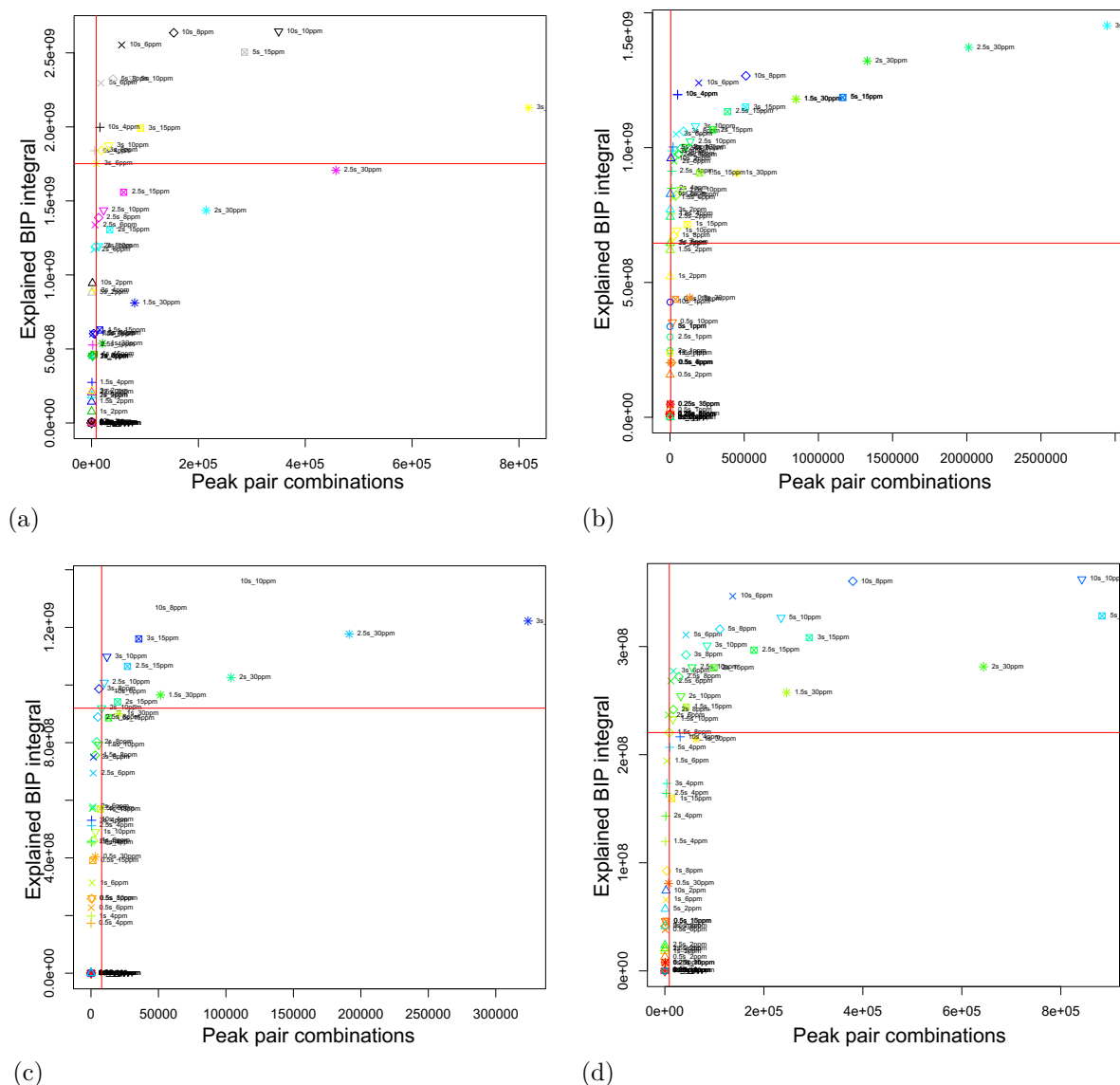


Figure IV.12.: Different rtE and $ppmE$ settings for the four datasets:

Figure IV.12a ESI+ HILIC ($rtE = 3s$, $ppmE = 6$).

Figure IV.12c ESI- HILIC ($rtE = 2s$, $ppmE = 10$).

Figure IV.12b ESI+ RP ($rtE = 1s$, $ppmE = 4$).

Figure IV.12d ESI- RP ($rtE = 1.5s$, $ppmE = 8$).

Explained BIP area (red signal in figure IV.11) is plotted against the total number of C-N-S combinations at a certain threshold. The same $ppmE$ and rtE values share the same shape and colour, respectively. Red lines mark the level of the as ideal determined $ppmE/rtE$ threshold.

Chromatography		HILIC	HILIC	RP	RP
ESI polarity		+	-	+	-
Detected peaks	None	402195	321285	283172	220553
	C	363811	274112	263984	194935
	N	367117	254474	252549	180861
	S	379594	246736	265491	177409
	CN	364628	250182	251459	152268
	CNS	362284	241506	250764	139291
None - X Pairs	C	1661761	1367959	351962	790943
	N	1627267	1311641	339283	752872
	S	1870273	1308926	448623	905909
None - X Pairs after elemental ratio	C	176550	118785	47273	63290
	N	147716	99156	46762	54052
	S	106747	65539	51721	46705
Double Constraint	C+N = CN	35215	7975	24018	9411
Triple Constraint	CN+S = CNS	16322	7876	10388	12408
High Quality Pairs	CNS	488	421	473	268
Reduction [%]		99.879	99.869	99.834	99.878

Table IV.8.: **Reduction algorithms's performance benchmark:**

This table shows the number of data points at several stages of the reduction algorithm for the four different datasets.

the algorithm, but rather the decision to exclude multiply charged molecules, such as proteins or (*E. coli* cell wall) carbohydrates, from the analysis. All of these multiply charged entities were identified by the charge indicator. The efficiency of the multiple labels shows that their use is not only necessary for constraining the total formula space, but also for drastically cutting down random combinatorial peak pairs, when using such an automated analysis.

It is very interesting to see that the numbers of true biological peaks are quite similar, for two completely different chromatographic modes (RP and HILIC) and positive and negative ionisation modes. This shows that the algorithm is highly effective, regardless of the chosen separation stage. The small number (< 1000) of biological peaks is perhaps disappointing, but agrees with other studies [31].

In the final step, the remaining features are used to see how many molecular compositions,

i.e. actual metabolites, can be generated.

IV.3.4.2. Testing the annotation by the formula generator

Before it was used to annotate the biological features, the formula generator was tested to see whether it could generate the unambiguous formula for a compound at 3 ppm accuracy, as predicted by the *in silico* experiments in section IV.3.1. To this end, 1000 random molecules were sampled from the PubChem database. Carbon, nitrogen and sulphur atom-counts as well as the monoisotopic mass (as denoted in the PubChem library) were passed onto the formula generator, to complete the hydrogen, oxygen and phosphorus counts. In all cases, those were correct: each time one and only one formula was generated. The average error between the database mass and the one generated by the formula generator was 2.3×10^{-5} ppm with a standard deviation of 7.2×10^{-4} ppm. These calculations, however, were performed without taking adducts into account, as this would require assumptions about the adduct distributions.

Chromatography	HILIC	HILIC	RP	RP
ESI polarity	+	-	+	-
High Quality Pairs	488	421	473	268
Possible compositions	2706	2566	1432	1685
After Isotope check	895	833	491	657
After Lewis check	676	817	383	644

Table IV.9.: **Number of elemental compositions calculated for the test datasets:**

The total number of possible compositions, under consideration of adducts (table C3.3), is calculated from the reduced datasets (table IV.8). It is then checked whether the adduct-formula combination is in agreement with the corresponding m/z values of the different labelling regimes. All compositions are then subject to a Lewis octet rule check [344].

Table IV.9 shows the number of elemental compositions calculated for the four datasets. In contrast to predictions, more than one possible formula per isotope pair was initially calculated. This is due to two reasons. The empirically determined ideal ppm thresholds had to be chosen slightly higher than the 3.7 ppm specified in the theoretical analysis in section IV.3.1. Indeed, the dataset with the ppm cutoff closest to 3.7 ppm (RP+ with 4ppm) actually yielded fewer compositions than initial pairs, after all checks. However,

another, possibly more significant reason, is the inclusion of adducts- 14 adducts for negative and 31 adducts for positive ionisation mode. Thus, every isotope pair results in a several fold higher number of adduct/formula pairs that require to be resolved - something not taken into account during the theoretical analysis. However, two filters were used to reduce the number of adduct/formula pairs. It proved to be very effective to compare whether those pairs were also supported by the other isotope label regimes. This reduced the number of possible adduct/formulae pairs by two to three fold. Another filter is the Lewis octet rule check [344], *i.e.* whether the molecular compositions have no unpaired electrons. Molecules with unpaired electrons are radicals. Compared to other ionisation processes, such as electron impact ionisation, radicals are unlikely to be formed during the electrospray ionisation [619].

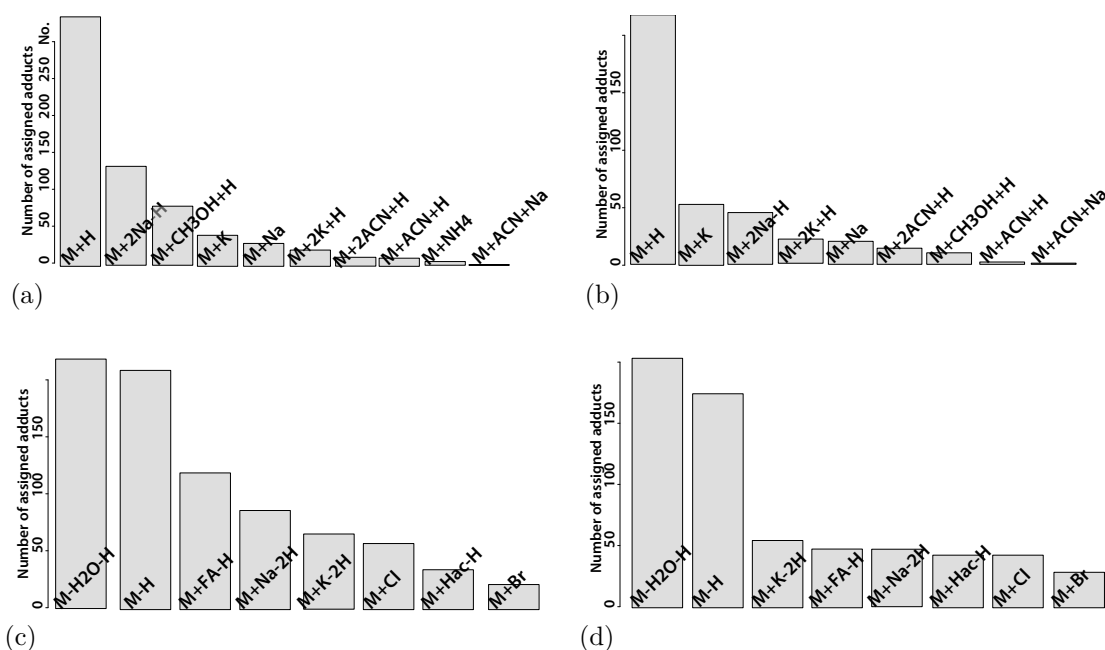


Figure IV.13.: **Adduct distributions in the test data sets:**

Adducts assigned for the four datasets. HILIC+ (IV.13a), RR+ (IV.13b), HILIC- (IV.13c) and HILIC- (IV.13d)

Figure IV.13 shows the post-filter list of adducts assigned during the formula generation step. Many fewer adducts than theoretically possible were assigned, demonstrating the high specificity of the assignment and filters. Indeed, the assigned adducts are highly plausible. As expected, the M+H and M-H ions were (nearly) always the most abundant,

followed by alkali and modifier salt adducts. Also, the charge indicator and filters seem to have been effective, as not a single multiply charged ion adduct was assigned. In negative mode, the water loss (-19 m/z) was slightly more frequent than the (-1 m/z) M-H adduct, although (when compared to related studies) it should be much more rare. In nearly all instances, where an adduct/formula pair with water loss (M- H_2O -H) was assigned, a formula with M-H was also assigned, but with loss of $-H_2O$. This makes sense but cannot be circumvented, as the formula generator cannot discriminate between a H_2O lighter formula and a water loss. Thus, if one had to decide between the M-H adduct and H_2O loss, M-H may be a more probable choice. The same problem could apply to any adduct, which is caused by loss or aggregation of CHNOPS elements. Even though less adducts than theoretically allowed were detected by the algorithm, figure IV.13d clearly shows that using just the M+/-H adduct or only 3 different adducts per mode, as done by all other studies [234, 196, 31, 194], is not enough and will lead to misassignments or unassigned peaks.

To examine the molecular formulae assigned to the isotope pairs, rt-mz heat maps of the atom-counts were produced (HILIC ESI+ IV.14, reversed phase ESI+ (IV.15), HILIC ESI- IV.16, reversed phase ESI- (IV.17)). For formulae with multiple adduct assignments, the M+H, M-H ions were preferred. In the few cases of multiple formulae assignments, the one with the lowest ppm values (across all labelling regimes) was chosen.

Firstly, the high efficiency of the algorithm should be noted. It identifies non-redundant biological peaks from a huge background of potential detected peaks. Even though those features continue beyond 1 kDa, the metabolite candidates are almost exclusively below that value. This shows that the "untargeted analysis of biomolecules below one kDa" [246] is a fairly good definition of metabolomics. It can also be seen that similar atom-counts cluster, rather than being randomly arranged across the $RT \times m/z$ domain. This is true for the atom-counts of the directly determined elements (C/N/S), but also for the calculated ones (H/O/P), demonstrating that the data extraction and formula generation algorithm both deliver plausible results. Indeed, the incidents of extreme atom-counts, such as the occurrence of 10+ phosphorus atoms per molecule are very rare - albeit not implausible, as element probability ratios were used during molecular formula generation. The bulk of the molecules, though, are highly plausible.

Both chromatographic modes showed their expected behaviours. HILIC had very few

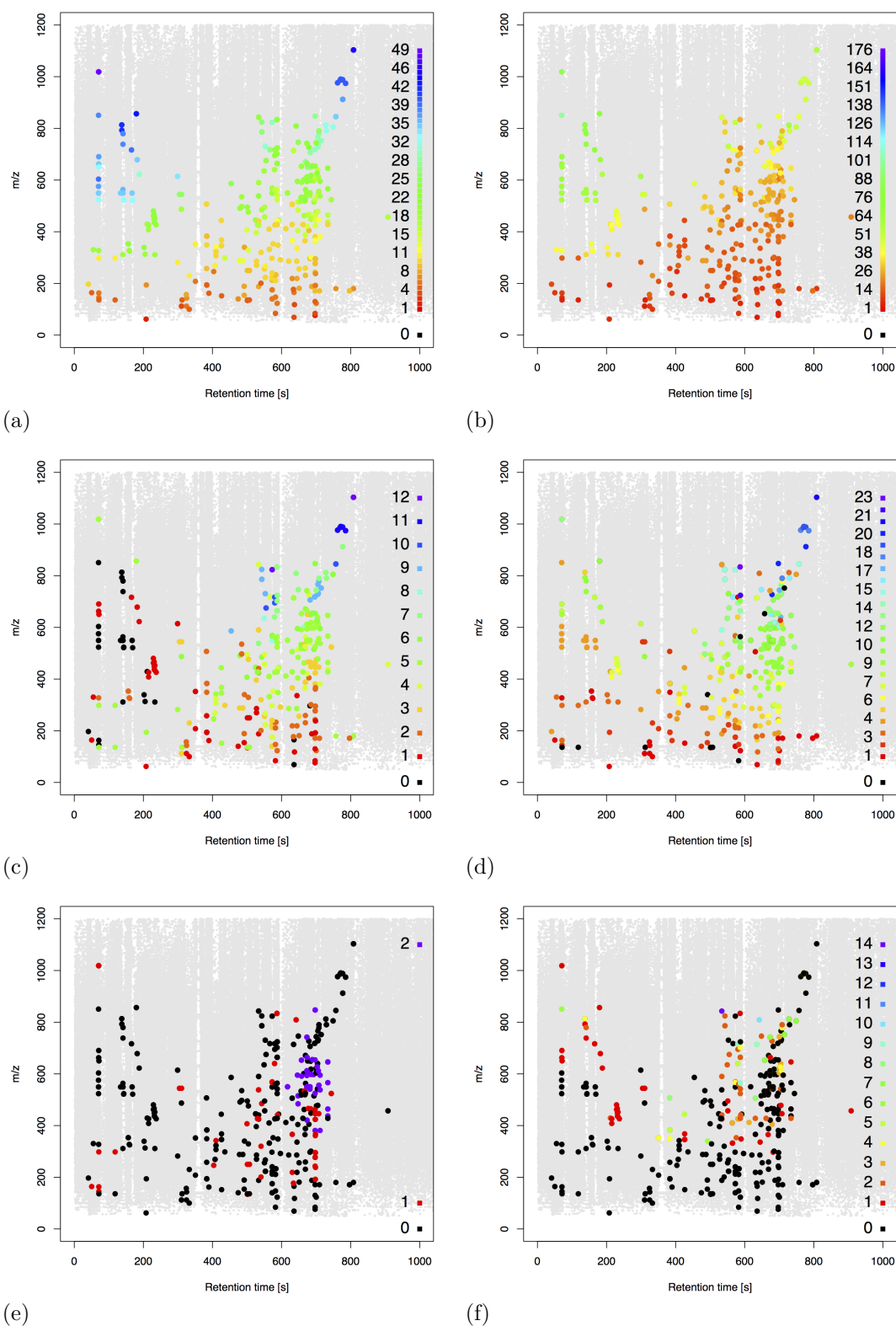


Figure IV.14.: Molecular formulae of HILIC ESI+:

RT vz m/z plots with atom-counts for C (IV.14a), H (IV.14b), N (IV.14c), O (IV.14d), P (IV.14f), S (IV.14e). All (non-reduced) detected peaks are in grey.

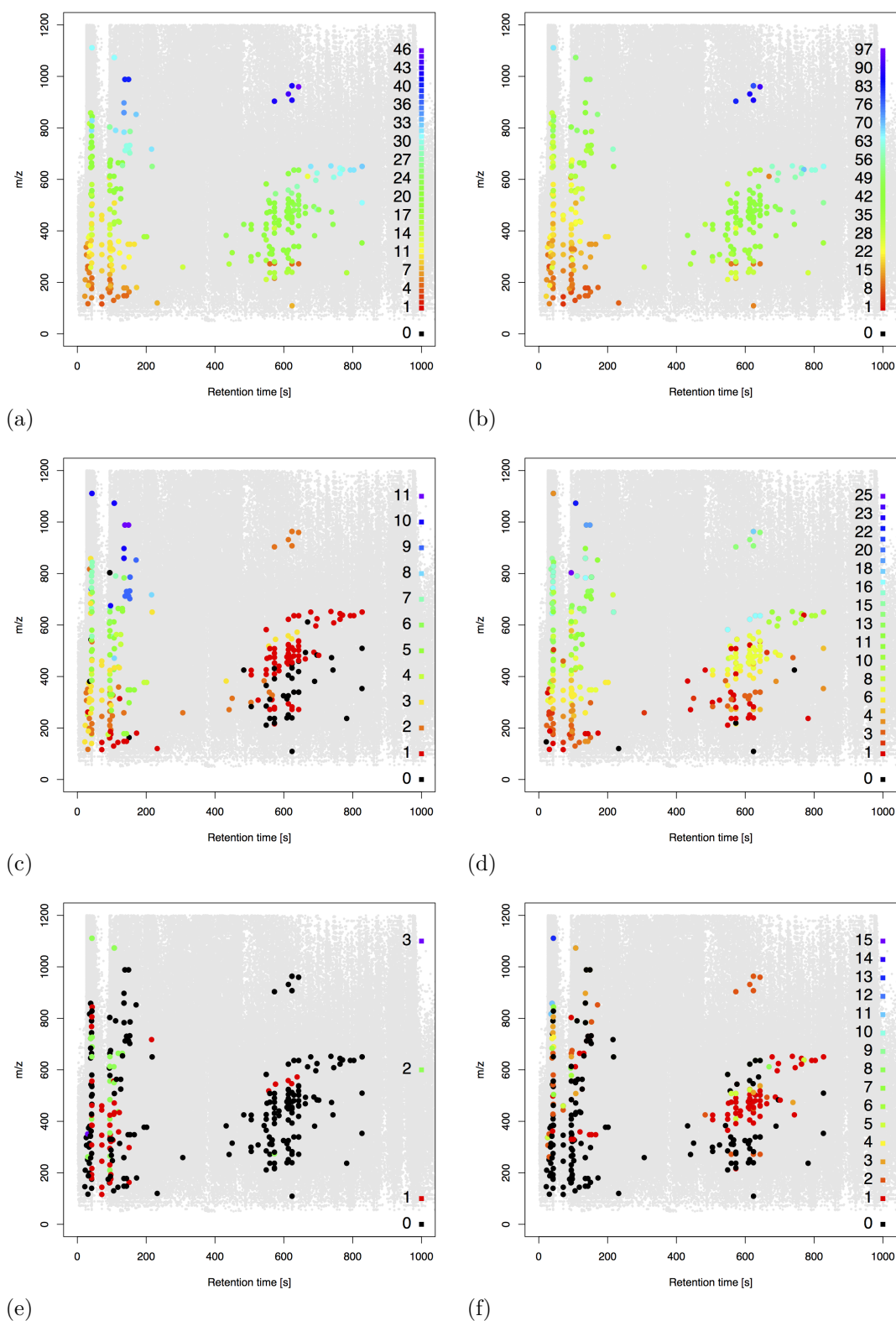


Figure IV.15.: Molecular formulae of reversed phase ESI+:

RT vz m/z plots with atom-counts for C (IV.15a), H (IV.15b), N (IV.15c), O (IV.15d), P (IV.15f), S (IV.15e). All (non-reduced) detected peaks are in grey.

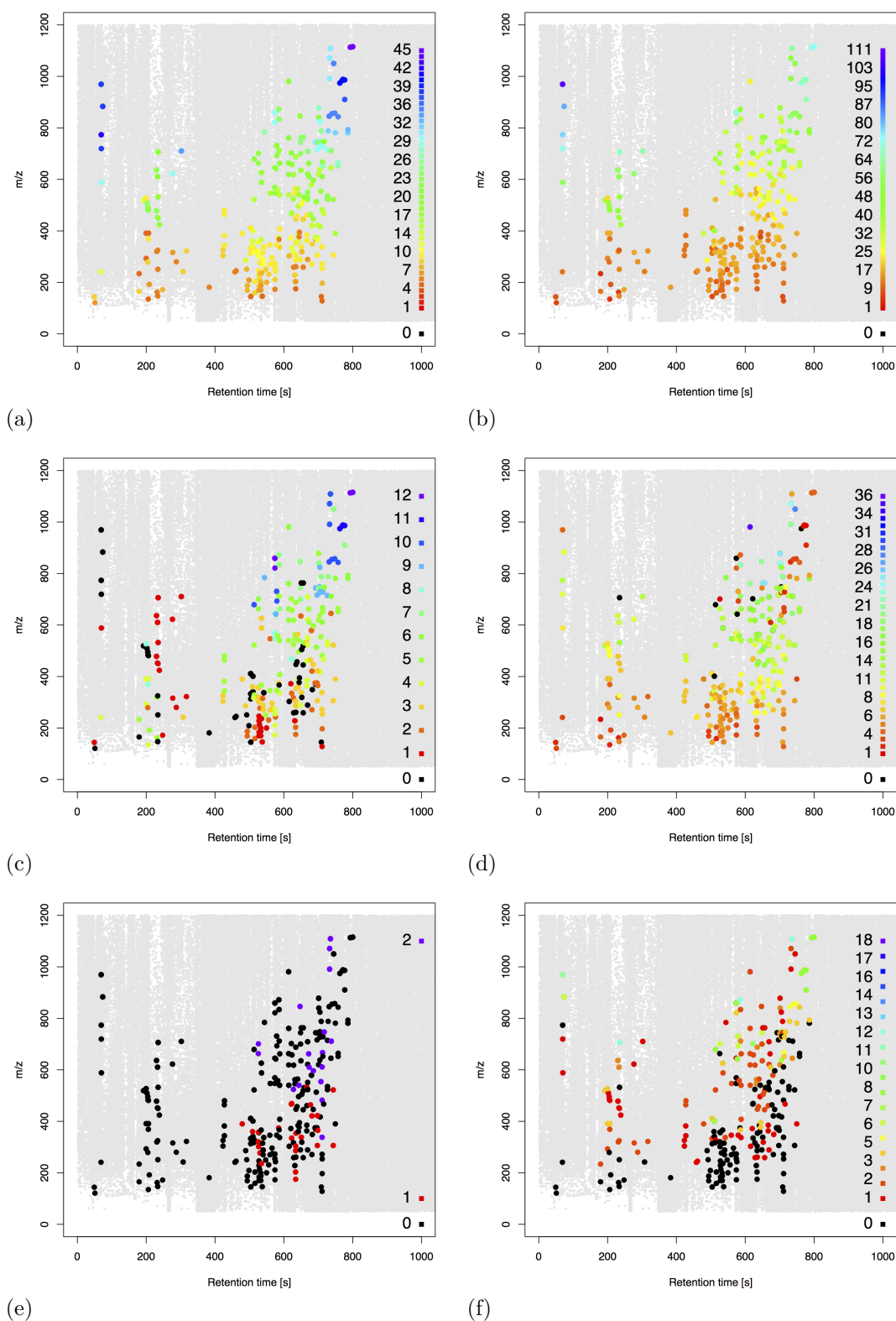


Figure IV.16.: Molecular formulae of HILIC ESI-:

RT vs m/z plots with atom-counts for C (IV.16a), H (IV.16b), N (IV.16c), O (IV.16d), P (IV.16f), S (IV.16e). All (non-reduced) detected peaks are in grey.

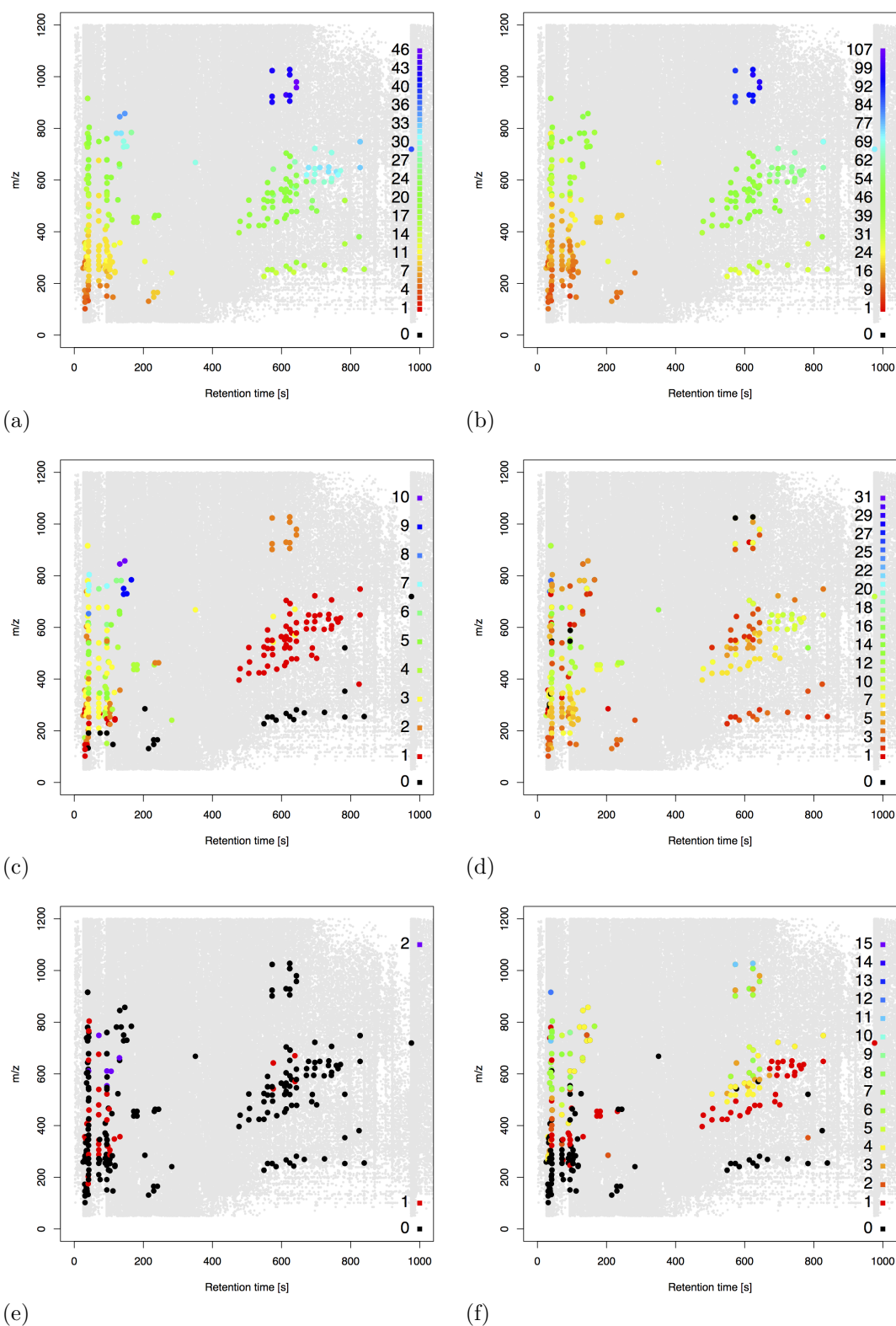


Figure IV.17.: Molecular formulae of reversed phase ESI-:

RT vz m/z plots with atom-counts for C (IV.17a), H (IV.17b), N (IV.17c), O (IV.17d), P (IV.17f), S (IV.17e). All (non-reduced) detected peaks are in grey.

void volume eluting molecules, compared to reversed phase. In HILIC, small (low carbon count) molecules eluted throughout the whole run, until late. In reversed phase, only carbon counts above 15 eluted after 10 min, showing many lipids, including phospholipids, as can be clearly seen from the nitrogen and phosphorus atom-counts. In HILIC, most sulphur containing compounds eluted late, *i.e.* they were rather polar, compared to reversed phase where they eluted early. Interestingly, the molecules eluting at the void volume in RP are a mix of low and high molecular weight compounds. However, most of the heavier ones also showed signs of high polarity, such as high nitrogen counts, particularly in ESI+, thus explaining their early elution despite high molecular weight.

When comparing positive and negative modes, compounds that have the same atom-counts, retention time and similar molecular weight, *i.e.* the same compound ionising in both modes, can be found. However, plenty of compounds are unique to a single mode, highlighting the value of running both modes.

It is not useful to compare the number of generated formulae with those reported in other isotope based annotation literature, as most of them use *Arabidopsis thaliana* samples and a different analytical setup. Baran *et al.*[31] though, used carbon and nitrogen constraints to annotate the metabolites of the cyanobacterium *Synechococcus sp.* PCC 7002 by reversed phase chromatography. They were able to generate 82 unique compositions, of which they could confirm 58 by fragmentation studies. 26 of the compositions though, were neither contained in KEGG [450], nor the MetaCyc [476] database. Even more relevant is the work of van der Werf *et al.*[588]. They tried to assign as many compounds as possible from *E. coli* extracts, on GC-MS and LC-MS, using authentic standards. Overall, they were able to assign 176 metabolites, leaving 255 unknowns. In reversed phase, they were able to annotate 21 out of the 50 detected peaks, and in HILIC, 37 of the 87 compounds. This is far less than the number of peaks with empirical formulae annotated in the four *E. coli* datasets of this chapter.

As this thesis is more concerned with *C. elegans* rather than its food source, *E. coli*, no in-depth analyses of the molecular compositions, *e.g.* database queries, were conducted. The *E. coli* data were for algorithm test purposes only, as the corresponding *C. elegans* datasets were of insufficient quality. Nevertheless, some ad hoc checks on individual molecular compositions were performed. Several database matches were found to metabolites that

are highly specific to bacteria and *E. coli* in particular. Those include enterobactin (RP-; 350 s; 668.14 Da) and deoxythymidine diphosphate (dTDP)-l-rhamnose (RP-; 91 s; 547.07 Da), which are *E. coli* iron scavengers (siderophor) [449] and precursors of *E. coli* virulence factors [436, 138], respectively.

IV.4. Conclusion

Two major current challenges of untargeted LC-MS metabolite analysis were addressed in this chapter, both via stable isotope enrichment of the sample: reduction of large datasets to non-redundant, biologically relevant features, and the annotation of their molecular composition.

Simulations showed that by constraining the carbon, nitrogen and sulphur-count, unambiguous formulae up to 1 kDa can be calculated *ab initio* with current ToF mass accuracy. The final, fully automated algorithm was able to reduce datasets by over 99.9 % and generated highly plausible molecular compositions for the remaining 0.1 % peaks.

This method can be used for many applications. An exhaustive metabolite characterisation of a new biological matrix, if it can be isotope enriched, is the most obvious example. However, it may also be useful for samples that cannot be isotopically labelled, such as urine or plasma. As a large subset of (central carbon and lipid) metabolites is shared between species, a characterised extract could be spiked into urine or plasma, thus providing a cheap and comprehensive standard mixture of isotopically labelled standards for quantification, including many compounds which are not commercially available. This method is also useful for analytical method development. By determining which features are of biological origin, a chromatographic method can be optimised to resolve these, rather than a TIC/BIP full of unknown features, as is often the case.

The approach developed here will be used in the next chapter, the *C. elegans* longevity-mutant screen. Metabolomics is usually defined as completely untargeted profiling. However, as shown, most information acquired and compared via means of multivariate analyses, is redundant and not even biological. Hence, using only peaks identified to be of biological origin should allow for better data handling and improved statistics, yielding higher quality, and most importantly, relevant models.

V.. A metabolic model of longevity in *C. elegans*

Aims and Ojectives:

- Find and annotate *C. elegans* metabolites using the workflow from chapter IV.
- Match *C. elegans* metabolites to longevity-mutant profiling experiment
- Find a common metabolic signature of longevity

V.1. Introduction

V.1.1. Motivation

This thesis' aim is to develop analytical and computational approaches that help understand the metabolic basis of pathways that determine life expectancy in the model organism *C. elegans*.

In chapter III.4, a dataset comparing mitochondrial mutations, which lead to increased or shortened life expectancy, was analysed using a newly-developed GC-MS workflow. However, this dataset contained only two different life expectancy extending mutations from one single mode of lifespan extension. So far, using whole genome RNAi screens [224, 227, 439], over 400 [415, 651] *C. elegans* genes have been discovered that modulate the rate of ageing and lead to a long-lived phenotype. These genes belong to various different signalling pathways, with a large proportion of the genes and pathways having human homologues [522]. Furthermore, many of these pathways are closely tied to an organism's core metabolic functions, such as lipogenesis or carbohydrate and amino acid metabolism. Two studies (published while the experiments described in this thesis were ongoing) have reported distinct metabolic signatures for two of the major pathways leading to lifespan extension in *C. elegans*: the insulin like signalling (ILS) pathway [183] and the mitochondrial respiration pathway [75], which was also analysed in chapter III.4.

However, many more pathways involved in lifespan regulation remain to be investigated. There has been surprisingly little experimental effort to compare all these different pathways to yield a more holistic picture - as might have been expected in the age of systems biology, especially given that more and more evidence emerges how many of those life-extending signalling pathways are either directly co-regulated, linked by core metabolic processes, or share common effector mechanisms [634, 437, 524, 653, 255]. Therefore, in this chapter, the metabolic signature of a diverse array of 24 different long-lived mutant strains will be determined. This will allow the creation of cross-pathway models, to determine global factors related to ageing and longevity in *C. elegans*.

The samples will be analysed by latest generation high resolution LC-MS instrumentation. As shown in the last chapter (IV.), the major share of peaks in LC-MS metabolomics data may actually not be biologically relevant at all, or else redundant in some way (isotope or

adduct peaks). Therefore, the mutant screen will be analysed in a semi-targeted manner, with only features identified to be of biological origin used for analysis. Annotating peaks with molecular compositions may yield chemical or even metabolic pathway information about a feature, allowing for greater biological understanding.

V.1.2. Selection of long-lived mutant strains

Firstly, the genes to be analysed in the mutant screen need to be selected. In principle, there are three major levers to modulate life expectancy and the ageing process in *C. elegans*: genetic, pharmacological or environmental [456, 107, 210, 488]. Genetic intervention can either be permanent, by introduction of a loss of function mutation in an allele, or temporary, by silencing a gene using RNAi [294]. Environmental interventions, however, are difficult to control [385], with non-optimised conditions often yielding negative results. Furthermore, they are rarely specific, impacting various, often unknown, physiological functions, which would make comparison of different interventions difficult. Pharmacological interventions may also give a complex response, and also depend on dose and bioavailability is non-trivial for worms as many drugs have very poor cuticular penetration.

Gene silencing by RNAi can easily be performed by feeding bacteria which express the required siRNA sequence. An advantage of using RNAi is that the time of silencing can be timed. Thus, life-extending genes that are essential for development [121], such as target of rapamycin (TOR), can be knocked down in adult worms. In order to confirm a gene's successful knockdown, either its expression or loss of function would have to be probed, using molecular methods, such as gel electrophoresis and PCR, or a longevity assay, respectively. However, both methods were not feasible within the scope and resources of this project.

For those reasons, it was decided to focus on mutants. A list of all selected strains is given in figure V.1. The criteria for inclusion were that the mutant strains had to be available from the CGC, and must have lifespan extension documented in a peer-reviewed publication. The list contains extremely well characterized genes (*e.g.* *daf-2*), but also some which have been much less investigated. This provides a chance to make use of metabolomics as a tool for functional genomics [485] and potentially shed more light into a gene's mode of action. For some pathways, knockouts of different proteins of the same pathway were

Strain	Gene(Allele)	Group	Extension [%]	DAF-16 relationship	Literature
BA843	<i>spe-26(it118)</i>	Fertility	45	dependent	[431]
CB4037	<i>glp-1(e2141)Ê</i>	Fertility	75	dependent	[21]
CB1370	<i>daf-2(e1370)</i>	ILS	100	upstream	[310]
BQ1	<i>akt-1(mg306)</i>	ILS	48	upstream	[463]
TJ1052	<i>age-1(hx546)</i>	ILS	100	upstream	[423]
CB4876	<i>clk-1(e2519)</i>	Mitochondrial	40	independent	[636]
MQ989	<i>isp-1(ctb-1)</i>	Mitochondrial	60	independent	[169]
MQ887	<i>isp-1(qm150)</i>	Mitochondrial	65	independent	[169]
MQ1333	<i>nuo-6(qm200)</i>	Mitochondrial	70	independent	[649]
MQ513	<i>daf-2(e1370) clk-1(e2519)</i>	Mitochondrial + ILS	130	mixed	[341]
KB6	<i>rle-1(tm2447)</i>	DAF-degradation	25	DAF-degradation	[362]
KX15	<i>ife-2(ok306)</i>	Protein Turnover (TOR)	25	dependent	[228]
RB1206	<i>rsk-1(sv31)</i>	Protein Turnover (TOR)	25	independent	[228]
YH461	<i>ifta-2(tm1724)</i>	Sensory	30	upstream part.	[516]
FK229	<i>pkg-1(ks61)</i>	Sensory	25	upstream	[243]
PR801	<i>che-3(p801)Ê</i>	Sensory	100	partially dependent	[19]
PR802	<i>osm-3(p802)</i>	Sensory	65	partially dependent	[19]
PR813	<i>osm-5(p813)Ê</i>	Sensory	120	partially dependent	[19]
PR678	<i>tax-4(p678)Ê</i>	Sensory	90	partially dependent	[19]
PR811	<i>osm-6(p811)Ê</i>	Sensory	60	partially dependent	[19]
CB246	<i>unc-64(e246)Ê</i>	Sensory	35	upstream	[4]
DA1116	<i>eat-2(ad1116) Ê</i>	Dietary Restriction	40	independent	[342]
ZG31	<i>hif-1(ia4)</i>	HIF-1	25	upstream	[92]
LG394	gls3 [<i>sir-2.1(+)</i> + <i>rol-6(su1006)</i>]	Sirtuins	16	upstream	[602]
LG398	gls101 [<i>rol-6(su1006)</i>]	CTRL	-	-	[602]
HT941	lpIn1 [<i>jnk-1(+)</i> + <i>rol-6(su1006)</i>] <i>Ê</i>	c-Jun N- kinase	25	upstream	[454]
RB1072	<i>sod-2(ok1030)</i>	ROS	50	???	[589]

Table V.1.: *C. elegans* strains used for the longevity mutant screen:

This table includes the strain name and the targetted allele. Furthermore the "mode of action", the observed lifespan extension (compared to N2 wild type) reported in the literature, and the relationship to the transcription factor DAF-16 is given.

selected. The function of all the selected genes will be discussed in the next section, which reviews the different mechanisms of genetic control of ageing in *C. elegans*.

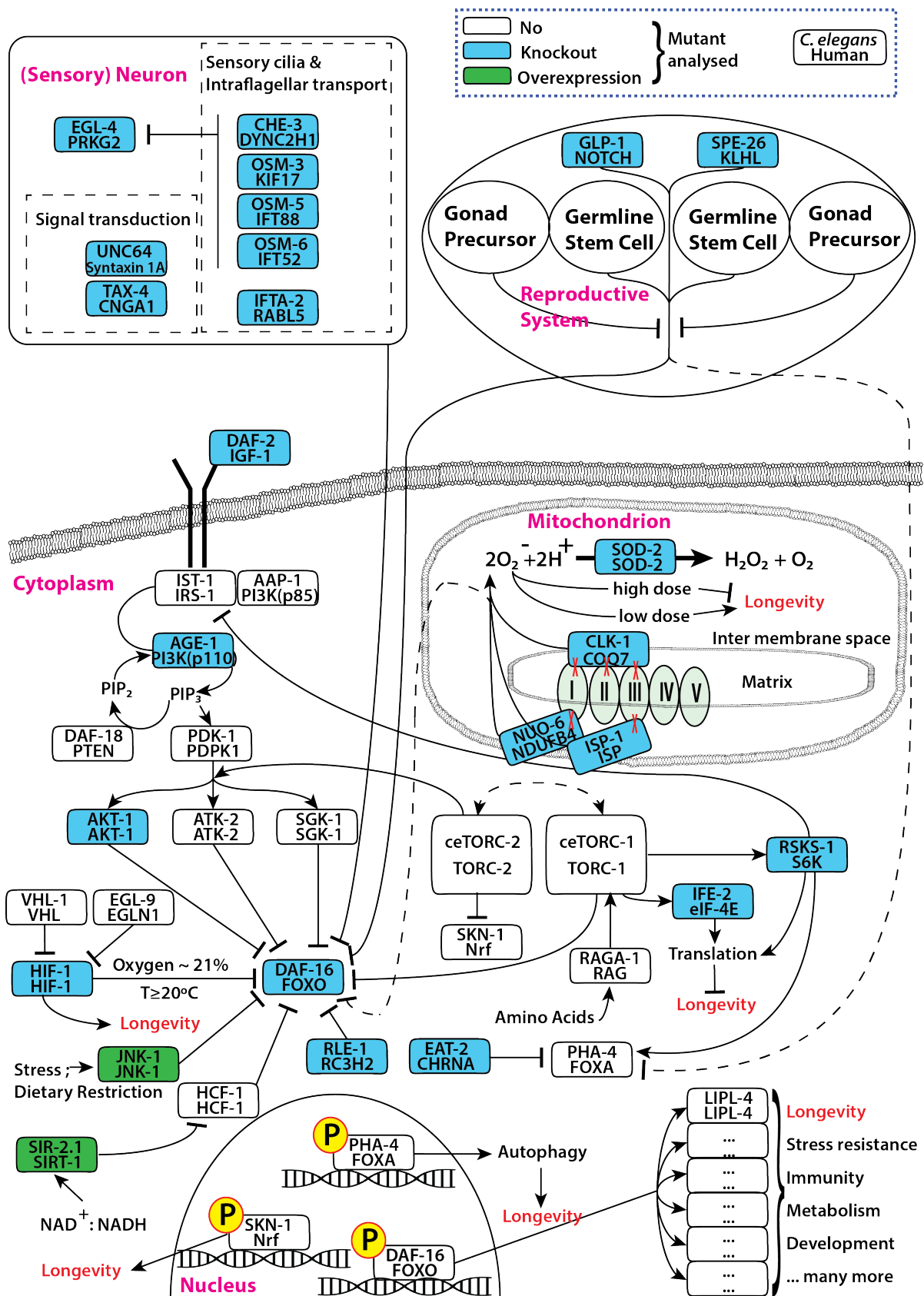
V.1.3. Longevity promoting genes in *C. elegans*

This section discusses different mechanisms which control ageing and life expectancy in *C. elegans*, with respect to the genes that will be analysed in the mutant screen. If possible, they will be classified into categories. Also, if known, links between the regulatory gene network and the metabolites themselves will be discussed.

V.1.3.1. Insulin like signalling (ILS)

The first discovered [182, 310] and best characterised pathway that governs lifespan in *C. elegans* is the insulin like signalling (ILS) pathway, shown in figure V.1. Just like in humans, insulin secretion is triggered upon the presence of food. Unlike humans, *C. elegans* expresses over 40 different insulin-like peptides, which may agonize or antagonize the insulin-like growth factor 1 (IGF-1) receptor [480], more generally known as the DAF-2 receptor in the ageing literature [317]. Upon activation, this transmembrane receptor tyrosine kinase triggers a complex phosphorylation signalling cascade, which ultimately inhibits DAF-16; a main hub of longevity [653]. This signalling cascade works via the insulin receptor substrate IST-1 [635] and AGE-1, the *C. elegans* ortholog of a phosphoinositide 3-kinase (PI3K) catalytic subunit [182], with AAP-1 as adaptor protein [635]. AGE-1 converts the second messenger PIP_2 (phosphatidylinositol biphosphate) to PIP_3 (phosphatidylinositol triphosphate) [423], which results in the activation of PDK-1, a 3-phosphoinositide-dependent kinase [462]. The phosphatase and tensin homolog (PTEN) catalyses the reverse reaction [452]. PDK-1 activates the serine/threonine kinases AKT-1, AKT-2 [463] and SGK-1 [241] by phosphorylation. Each of them, independently, may phosphorylate the forkhead box (FOXO) transcription factor, DAF-16 [368, 453], which then becomes inactive and shuttles from the nucleus to the cytosol, with the help of 14-3-3 proteins [81].

If the activity of DAF-2 or any of the proteins in the ILS cascade is reduced, either by KO mutation or RNAi knockdown, DAF-16 returns to the nucleus and becomes active.

Figure V.1.: Pathways of lifespan regulation in *C. elegans*:

All pathway data has been collated from literature. The *C. elegans* protein and the corresponding human homologue/orthologue are given [568, 522, 654]. For details, please refer to section V.1.3

It transcribes a whole array of genes [432, 354, 400, 519], which lead to a long-lived phenotype [238]. Depending on the mutated allele and severity of mutation, and therefore residual ILS activity, the lifespan extension may be up to 100 %, effectively doubling the worm's lifespan. This phenotype, however, is completely suppressed by DAF-16 knockout, indicating that DAF-16-modulated gene expression is wholly responsible for the ILS lifespan extension: the downstream genes of DAF-16 are key players in lifespan regulation, development, stress resistance, reproduction, and metabolism [432, 354, 400, 519].

Like any other protein, DAF-16 undergoes proteasomal degradation. RLE-1 is an E3 ubiquitin ligase, which catalyses DAF-16 polyubiquitination and subsequent degradation. Unsurprisingly, if RLE-1 function and therefore degradation is impaired, worms live longer [362].

As the ILS pathway is highly conserved, genes involved in the ILS cascade, have homologues in *D. rosophila* and even mammals, like mice or humans [437]. In drosophila [102, 565] and mice [247, 53] knockout of the IGF-1 receptor or downstream components also leads to an increased lifespan. Demographic studies in humans have connected single nucleotide polymorphisms (SNP) in FOXO3A, the human homologue of DAF-16 [653], to human longevity [662].

One of the DAF-16 downstream targets is the triglyceride lipase LIPL-4, which cleaves triglycerides. It was shown that upon disruption of the ILS pathway LIPL-4 is overexpressed and promotes DAF-16's longevity phenotype [608]. DAF-2 knockout worms even showed a decreased total triglyceride content. Recently, it was discovered that lipolysis works via mechanisms of autophagy [526], including in *C. elegans* [458]. Indeed, LIPL-4 based lifespan extension requires the induction of autophagy [345], which is mediated by the TOR pathway.

V.1.3.2. Protein Turnover, autophagy and dietary restriction mediated by target of rapamycin (TOR)

Another pathway, which is closely linked to insulin like signalling, is the Target of Rapamycin (TOR) pathway [298, 348]. TOR divides into two protein complexes, with distinct functions: TOR complex 1 (TORC1) and TOR complex 2 (TORC2). TORC1 is

a key hub for growth and nutrient sensing, integration of these signals and initiation of many response mechanisms, centred around cell growth. For example, it can sense amino acid concentrations via the RAG GTPase [508]. TORC2, which has been characterised to a far lesser extent, is insensitive to nutrient status, but regulates cell shape by signalling to the cytoskeleton. Over the last few years, TOR has also been identified as a major determinant of ageing [398] and age related disease [286]. Even though TOR, like ILS, is a very conserved pathway, there are species differences. Some key components of mammalian or even drosophila TOR do not have *C. elegans* homologues [286]. Thus, care must be taken when comparing TOR-related studies performed with other species. Therefore, mechanisms explained in this section will be restricted to the current knowledge of the *C. elegans* TOR (ceTOR) signalling network.

In 2003 Vellai *et al.* [596] reported that reduction of ceTOR activity by post-larval RNAi knockdown (knockout mutations lead to developmental arrest) increased lifespan. Other studies followed [282, 228, 523, 250]. They generally knocked down LET-363, the common protein of ceTORC1 and ceTORC2. This led to inconsistent results regarding which of the two TOR complexes increased lifespan, and whether DAF-16 is required for this increase. Very recently, a study [497] selectively silenced Raptor (DAF-15) and Rictor (RICT-1), proteins specific to ceTORC1 and ceTORC2, respectively. They showed that ceTORC1 silencing leads to a DAF-16-dependent increase in longevity, by activating DAF-16 and its downstream genes. Longevity induced by reduction of ceTORC2 activity, however, was independent of DAF-16, but dependent on and mediated by SKN-1. SKN-1 is orthologous to the mammalian Nrf (Nuclear factor-erythroid-related factor). Nevertheless, ceTORC2 can still undergo cross-talk with DAF-16, by activating AKT1/2 and SGK-1.

Also highly interesting is the lifespan extension associated with ceTORC downstream targets, particularly for ceTORC1. One of ceTORC1's main effector mechanisms is modulating the rate of protein biosynthesis. This is done by direct activation of RSKS-1, a ribosomal protein S6 kinase (S6K) and IFE-2, a translation initiation factor. Knockout of either [557, 228], leads to an increased life-span, independent of DAF-16. RSKS-1 controls a series of downstream targets from different pathways, which also modulate ageing. By inhibiting the insulin receptor substrate, it even interferes with ILS. RSKS-1 also inhibits PHA-4, a FoxA transcription factor. In fact, RSKS-1 knockout dependent longevity even requires PHA-4, but knockdown of PHA-4 itself decreases lifespan. [523]

PHA-4 itself is also a crucial regulator of lifespan. A mutation in the EAT-2 protein, which corresponds to a ligand gated ion channel subunit, closely related to the non-alpha-subunits of nicotinic acetylcholine receptors, also increases lifespan [342]. This effect is PHA-4 dependent [461], and DAF-16 independent. The EAT-2 mutation disturbs pharyngeal pumping and limits food intake, therefore mimicking dietary restriction (DR). The effects of DR were first discovered in rats in the 1930s [397]: reducing caloric intake without malnutrition extends lifespan and decreases age-related diseases. This effect has been repeatedly reproduced in many organisms, such as yeast [369] and *drosophila* [481]. Increased long term survival has yet to be shown for humans. However, as very recently reported, it has a positive effect on ageing, as tissue samples of dietary restricted patients show transcription profiles similar to the ones of younger controls [411]. Furthermore, 6 years of DR can already significantly reduce the risk for atherosclerosis [179]. In *C. elegans* DR can be induced in various ways, such as dilution of the bacterial food source [385], intermittent feeding [250], complete removal of food [292], and drugs [478] or mutations which mimic DR, such as EAT-2. DR regimes in *C. elegans* extend lifespan by increasing autophagy [283, 225], which is mediated by downstream targets of PHA-4 [461].

Autophagy is a major catabolic mechanism, which degrades unnecessary and dysfunctional components of a cell. During autophagy, the phagophore expands into a double-membrane autophagosome. It surrounds cell components, such as proteins, lipid droplets [526] and even whole organelles, such as mitochondria, which are to be degraded. After encapsulation, the autophagosome fuses with a acidic lysosome filled with hydrolytic enzymes. After the resulting autolysosome contents are degraded, the basic building blocks, such as amino and fatty acids, are released into the cytoplasm [330, 7]. Thus, in times of scarce resources, such as dietary restriction, autophagy provides a way of recycling basic biomolecular building blocks and also providing energetic substrates. Unsurprisingly, not only DR requires autophagy for lifespan extension, but also longevity derived from ceTORC inhibition [225]. Interestingly, however, longevity induced by downregulation of protein biosynthesis, such as IFE-2 or RSKS-1 knockdown, does not require autophagy [460, 225]. Even more intriguingly, ILS-mediated longevity and the long term diapause of the dauer larva do require autophagy in addition to DAF-16 activation [410]. This shows that much remains to be elucidated regarding the exact interplay of longevity pathways and effector mechanisms.

V.1.3.3. Sirtuins (SIR)

Sirtuins are a family of highly conserved nicotinamide adenine dinucleotide (NAD⁺)-dependent deacetylases [270]. They were first discovered in yeast and shown to have an impact on ageing [291]. Shortly after, it was demonstrated that overexpression of SIR2 and SIR-2.1, in *C. elegans* [571] and *drosophila* [502], respectively, extends lifespan.

Recently, these findings have come under scrutiny. Several labs jointly showed that the apparent longevity effects of the *drosophila* study were due to improper choice of controls. The lifespan increase reported for *C. elegans* may be due to a second, unrelated mutation that was accidentally induced into the *C. elegans* SIR-2.1 overexpression mutant, together with the *sir-2.1* vector [73]. However, the two studies have been rescued by the use of proper controls [30] and a clean, outcrossed *sir-2.1* overexpression mutant [602]. Even though the clean strain is still long-lived, the absolute lifespan extension is less than initially reported. An eight times outcrossed *sir-2.1* overexpression strain will be used for the current study. However, *rol-6*, a collagen mutation that gives a dumpy phenotype, was used as transformation marker, *i.e.* to select transgenic lines where *sir-2.1* was successfully inserted. Therefore a *rol-6* mutant must be used as control, rather than the regular N2 wild type strain.

Recently, the mechanism by which SIR-2.1 extends lifespan in *C. elegans* was reported. SIR-2.1 inhibits host cell factor HCF-1, which itself inhibits DAF-16 [495]. Thus, by inhibiting the inhibitor, a higher SIR-2.1 activity activates DAF-16. SIR-2.1 is activated by a low NAD⁺ : NADH ratio, so it is a sensor of a cell's redox state. It is directly linked to citric cycle and mitochondrial (respiratory chain) activity, which also determine the rate of ageing, as will be discussed in section V.1.3.6.

Seven different sirtuins have been described in mammals, serving different functions, centred around various core metabolic and energetic pathways [90]. Very recently, a lifespan extension in Sirt6 overexpressing male mice was demonstrated [297]. This is exciting news, given that sirtuins can also be activated via drugs, such as resveratrol, a natural phenol found in red wine [257]. However, it is not yet clear whether these effects will be reproducible in humans [363, 658].

V.1.3.4. c-Jun N-terminal kinases (JNK)

c-Jun N-terminal kinases (JNK) are a highly conserved subfamily of mitogen-activated protein (MAP) kinases. They mediate cytoprotective responses to various forms of environmental stress, such as heat, inflammation, oxidative stress, ionising radiation and reactive oxygen species (ROS) [552, 631]. However, if overexpressed, JNK-1 increases *C. elegans* [454] and drosophila [607] lifespan. In *C. elegans*, this is promoted by DAF-16 activation, by phosphorylation of sites different from the ILS pathway activation sites [454].

Like *sir-2.1,ink-1*, which will be analysed in the mutant screen, is an overexpression mutant, that has the *rol-6* gene as a transformation marker. Thus, it will also use the *rol-6* mutant as control, rather than wild type.

No major studies on JNK lifespan extension in mammals have yet been reported.

V.1.3.5. Hypoxia-induced factor (HIF-1)

The hypoxia-induced factor HIF-1 is required for survival at hypoxic conditions. In low oxygen conditions, *C. elegans* has an increased lifespan [248], which is dependent on HIF-1 and DAF-16 [408, 357]. Until recently, the role of HIF-1 during normoxic conditions has been unclear, with four different groups reporting either increased and or decreased lifespan upon *hif-1* knockout [92, 664, 408, 429]. Leiser *et al.* [356] were able to resolve this ambiguity, by showing that *C. elegans hif-1*-related lifespan depends on temperature. At low temperatures (15°C), it is short lived, but at high temperatures (25°C), it has a significantly increased lifespan (compared to control). At the most commonly used temperature (20°C), which will also be used to analyse HIF-1 knockout mutants in this experiment, it is slightly long-lived. The increased lifespan is DAF-16 dependent [664, 356].

Interestingly, HIF-1 activation, for example by *hif-1* overexpression, also increases lifespan; this effect is even dose-dependent [664]. As discussed in III.4.3.2, *hif-1* activity can also be increased by deletion of the orthologue of the von Hippel-Landau tumor suppressor *vhl-1*, which is responsible for HIF-1 degradation. Longevity by HIF-1 activation, however, is DAF-16 independent. Instead, as discussed before (III.4.3.2), it seems to be mediated by the mitochondrial unfolded protein response (UPR^{mt}) [352] and therefore closely linked to mitochondrial mutations which increase longevity.

V.1.3.6. Mitochondria

Mitochondria produce a cell's energy by ATP regeneration through the electron transfer chain (ETC). The ETC are five protein complexes (I-V), with the first four building up a proton gradient across the inner membrane, which is then used to power an ATP synthase - complex V. RNAi screens have identified many ETC complex genes that increase lifespan when knocked down [351]. In fact, silencing of subunits from all complexes can extend lifespan, except for complex II - most probably because complex II also functions as the succinate dehydrogenase in the TCA cycle [261]. However, there a few key genes have been described in greater detail and will be included in the mutant screen.

clk-1 encodes a hydroxylase, required for the biosynthesis of ubiquinone (coenzyme Q, Q9) [341].

nuo-6 encodes for a subunit of the mitochondrial NADH dehydrogenase (ubiquinone) in complex I [649].

isp-1 expresses the Rieske iron sulphur protein (ISP), a subunit of complex III [169].

All these mutations also show slowed embryonic development and behaviour [636, 166], potentially supporting the "rate of living" theory of ageing [470]. This theory proposes an indirect correlation between metabolic rate and lifespan, *i.e.* the higher the metabolic rate, the lower the life expectancy. However, loss of a *ctb-1* (cytochrome b protein) allele suppresses the slow phenotype of the *isp-1* mutation, without compromising its lifespan extension. Thus, a slow metabolic rate, or reduced ATP generation, is not a necessary requirement for lifespan extension [137]. To examine uncoupling of these two effects, an *isp-1;ctb-1* double mutation will be analysed as well as the *isp-1* mutation.

There is increasing evidence that mitochondrial longevity is not mediated by simple slowing of physiological processes, but rather controlled by a sophisticated signalling network with different effector mechanisms. The exact mechanisms behind this, however, are far from being understood. The mode of lifespan extension induced by mutation, is different than the one by RNAi [649]. Mitochondrial mutations, but not RNAi, do not depend on au-

tophagy and other dietary restriction responses. There is also evidence that mitochondrial mutant longevity is somehow linked to ILS longevity, as the *daf-16;isp-1* double knockout is significantly shorter lived than *isp-1* itself, but still much longer than wild type [169]. Also both, *isp-1* [169] and *clk-1* [341] can increase *daf-2* KO lifespan, but this effect is not fully additive. Thus, as there seems to be a partial ILS involvement, a *daf-2;clk-1* double mutant will be analysed.

As already discussed in III.4.3.2, mitochondrial mutants require HIF-1 for lifespan extension [352] and also the induction of the mitochondrial endoplasmatic reticulum response [147, 92] (UPR^{mt}). In fact, in III.4.3.2 it was argued, that both could actually be linked via 2-hydroxyglutarate. However, Hwang *et al.*[268] argued that the retrograde longevity signal from the mitochondria could be in part mediated by reactive oxygen species (ROS).

V.1.3.7. Super oxide dismutase (SOD) and reactive oxygen species (ROS)

Reactive oxygen species (ROS) were proposed to be involved in ageing as early as 1956, by Denham Harman's "free radical theory of aging" [230]. Free radicals, such as superoxide (O_2^-), and other ROS, such as H_2O_2 , cause oxidative damage to biomolecules. They can lead to genetic mutation by DNA cross-linking, oxidation of polyunsaturated fatty acids in lipids (lipid peroxidation), oxidation of protein-bound amino acids (protein carbonylation) [538] and inactivation of enzymes, by coenzyme oxidation. According to this theory oxidative damage will eventually lead to diseases, such as cancer, and macroscopic effects, such as skin wrinkles. [174, 537]

In opposition to this, the long-lived mitochondrial mutants (*clk-1*, *isp-1*, *nuo-6*) [352, 648] and the insulin-like-signalling mutant *daf-2* [66] have increased levels of superoxide compared to wild type. Furthermore, several treatments that slightly increase superoxide levels also increase lifespan [518, 235]. In fact, increased superoxide levels are required for the lifespan extension, as the lifespan extension can actually be abrogated by the supplementation of ROS scavenging antioxidants, such as N-acetyl-cysteine or vitamin C [518, 648].

These observations have led to a new understanding of the role of ROS and support the theory of mitochondrial hormesis (mitohormesis) [494]. Hormesis describes the beneficial effect of a toxin in small doses, which in high doses is detrimental. This theory goes back to

the 19th century, to the Arndt-Schulz rule or Friedrich Nietzsche's dictum "That which does not kill us makes us stronger". Thus, small doses of ROS appear to be beneficial, whereas high doses lead to damage, accelerated ageing and premature death [189, 494, 236]. The molecular mechanisms, *i.e.* how ROS triggers signalling pathways, which lead to anti-ageing and longevity-ensuring effector mechanisms, are yet to be elucidated [268, 28].

Physiologically, the superoxide radical (O_2^-) is recycled by superoxide dismutase (SOD) enzymes. These metalloenzymes catalyse the reaction of superoxide (O_2^-) to molecular oxygen (O_2) and hydrogen peroxide (H_2O_2). The reactive peroxide is then converted into water and oxygen by enzymes such as catalase (CAT) and glutathione reductase (GPx). *C. elegans* has five SOD isoforms, with different expression levels, cytoplasmic or mitochondrial localisation and even different metals as coenzymes. Whether *sod* deletion or overexpression can lead to decreased or increased lifespan in *C. elegans* is still very controversial [139, 190]. However, it was recently shown that all five SODs can be knocked down without affecting normal lifespan, but increasing susceptibility to external stress [591]. VanRaamsdonk *et al.* [589] showed that knockout mutation of the mitochondrial manganese SOD-2, but not any other SOD, leads to increased lifespan. This result is in contrast to Doonan *et al.* [139], who did not report lifespan extension by any SOD knockout. To investigate how SOD is related to other lifespan extending genes, the *sod-2* KO strain used by VanRaamsdonk *et al.* will be analysed.

V.1.3.8. Fertility and germline cell signalling (GCS)

Removing *C. elegans*'s germ line stem cells, for example by micro laser beam surgery, increases their life expectancy by 60 %. However, when the whole gonad is removed, this effect is abolished [258]. In both cases, this leads to sterility. This is a strong argument against ageing theories, such as the life history theory [620], which see reproduction and lifespan as a tradeoff due to a limited amount of resources.

Germ-cell loss may also be mimicked by mutations which lead to a temperaturedependent reduction of progeny, and result in an increased lifespan. Two different mutations that lead to a germ-cell induced longevity, will be analysed. *spe-26* encodes a Kelch motif-containing protein, which is required for spermatogenesis [592]. Loss of function increases lifespan at 20°C [431]. Animals show a reduced fertility at 16°C, but zero fertility at 25°C. 20°C, however, as used in the current study, produced enough progeny for sample generation. *glp-1*

encodes for a transmembrane protein, which acts as a receptor that is required during mitotic germ cell proliferation. Its loss shows a similar behaviour with regards to reproductive capacity and lifespan as *spe-26* [21]. The lifespan extension of both mutants is DAF-16 dependent. However, double mutants have normal lifespan, despite still having reduced fecundity [431]. Again, this shows that the lifespan extension is more due to complex molecular signalling rather than a simple resource tradeoff [358]. Indeed, additionally to DAF-16, a whole signalling cascade centred around the steroid hormone receptor DAF-12 (a homologue of the human vitamin D receptor) is required [647]. Similar to ILS longevity, the expression of LIPL-4 is required for germ cell signalling (GCS) induced longevity [609, 345].

Even though it is tempting to think that ILS and GCS longevity responses simply converge at the level of DAF-16, this is not the whole story. Both effects are synergistic. DAF-2 suppressed worms which also have ablated germ cells are extremely long lived and healthy, with a more than 5 fold increased life expectancy [22]. Therefore, there may be a parallel mechanism which mediates longevity. Indeed, compared to ILS longevity, where autophagy seems to be optional, it is necessary for GCS longevity [345]. Thus, PHA-4/FOXA and possible other elements of the TOR signalling pathway are also required. Indeed, TOR knockdown does not increase *glp-1* mutant lifespan, suggesting a common effector mechanism and a (signalling based) link between reproduction and nutrient status.

V.1.3.9. Sensory perception and neuronal control

All pathways discussed above are triggered by biomolecules, such as nutrients, or physico-chemical conditions, like oxygen levels. Thus, they require sensory input. Over 100 years ago, Ivan Pavlov showed [469] that sensory cues and perception may suffice to trigger physiological mechanisms. In humans, for example, the anticipation of food, as mediated by the taste of a non-caloric sweetener, can trigger insulin release [575]. In *C. elegans* the odour of food (*E. coli*) can trigger ILS mechanisms and thereby influence lifespan [532]. Therefore, it is important to understand the mechanisms of how environmental cues are perceived and neuronally integrated to affect bimolecular pathways; with respect to ageing in particular [569].

Several genes encoding for sensory receptors [19] and neurons [8] increase lifespan when downregulated. Generally, compared to other genes discussed above, their signalling cas-

acades and effect on metabolism is poorly understood. No sensory mutants have yet been analysed using metabolomic approaches. *C. elegans* has 302 neurons, with 60 responsible for chemosensation and 60 for mechanosensation. For the actual sensing, those neurons express cilia, which are microtubule based subcellular organelles. They emanate from the cell surface and may have multiple sensory receptors. It is estimated that *C. elegans* has around 500 distinctive sensory receptors, most of them G protein-coupled receptors (GPCR), that trigger ion channels, such as cGMP-gated channels [280, 271, 370].

Many of the life-extending mutations are in proteins [19], which participate in the intraflagellar transport (IFT), a bisectonal transport system, along microtubuli. IFT is required for cilia formation and maintenance. The following mutations encode IFT components in cilia of chemosensory neurons. *osm-3* encodes for a kinesin motor protein. *osm-5* and *osm-6* encode proteins with human orthologues, required for morphology of chemosensory neurons and IFT. *che-3*: encodes a dynein heavy chain required for IFT [654]. All these mutants have deficient chemotaxis, and neither can they perceive odour. They also do not show the usual carbon dioxide avoidance of *C. elegans*[221].

ifta-2 encodes a homologue of the mammalian Rab-like 5 protein and is involved in IFT. However, compared to knockouts of other IFT proteins, cilia do assemble normally and worms do not show altered chemotactic behaviour, despite being long-lived [516].

tax-4 encodes a cyclic nucleotide-gated channel subunit, expressed in sensory neurons and is required for signal transduction. Its loss of function shows a very similar phenotype to the OSM mutants. [19]

egl-4 expresses a cyclic GMP-dependent protein kinase [243], which appears to act through the TGF-beta signalling pathway. Interestingly, *egl-4* is downstream of many *osm* and *che* genes and plays a role in various core physiological processes, such as dauer formation, growth and egg laying, during dietary restriction [560].

unc-64 encodes syntaxin, which is a plasma membrane receptor for intracellular vesicles. It is orthologous to vertebrate syntaxin and is expressed in the nervous system and secretory cells of *C. elegans*. It is essential for synaptic vesicle fusion and therefore signal propagation in the nervous system. *unc-64* is required for normal locomotion and possibly

also for insulin secretion [4, 654].

Interestingly, lifespan extension resulting from loss of function in all these genes is partially reversible by *daf-16* knockout. This demonstrates how tightly linked they are to common longevity assurance pathways. However, it can be anticipated that other pathways, such as TOR signalling, are also involved. [370]

Recently, a major *C. elegans* lab showed that chemosensory cilia function and IFT have a strong impact on lipid metabolism [386]. This is very interesting given all the recent evidence for the involvement of lipid metabolism in ageing [608, 347, 640, 1] and make RP-LC-MS metabolomics an ideal tool to gain more insights.

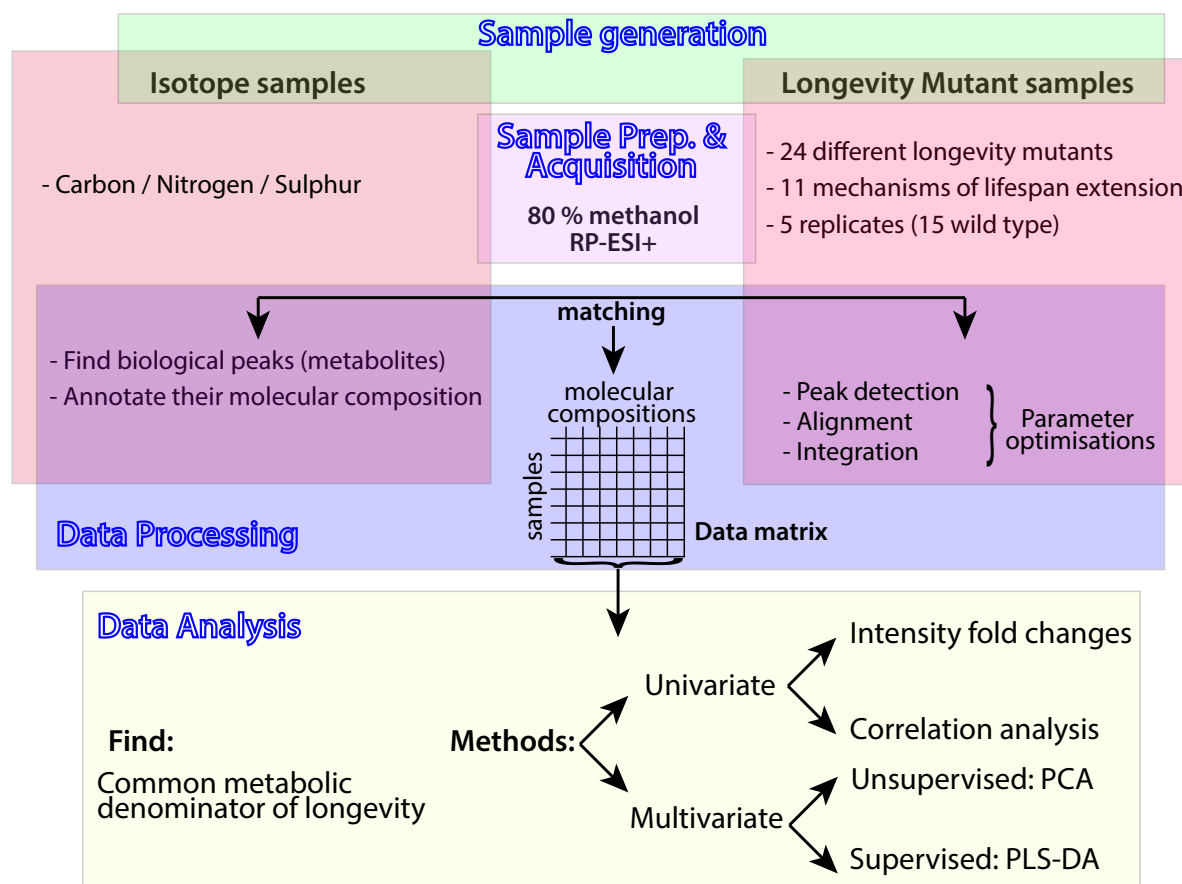
V.1.4. Outline of experiments

In this chapter, a large panel of long lived *C. elegans* mutants from various pathways/mechanisms of lifespan extension was analysed, to find out whether there is a common metabolic denominator of longevity.

Figure V.2 shows the general design of the study. In the first instance, two sample sets are created: carbon, nitrogen and sulphur stable isotope labeled *C. elegans* samples, for peak assignment, and the mutant samples. These included five biological replicates of 24 different mutants, from 11 different pathways of lifespan extension, wild-type worms had 15 replicates. All sample were extracted according to the consensus protocol from chapter II.: 80 % methanol extraction using a bead-beater. Both sample sets, the isotope samples and the mutant screen samples, were acquired in the same run by LC-MS, using a reversed phase separation and positive mode electrospray ionisation (ESI+).

Using the workflow and tools developed in the previous chapter (IV.), the isotope samples were used to identify all biologically synthesised metabolites in the *C. elegans* extracts and annotate them with their molecular compositions, providing a final data matrix for statistical analysis. The parameters for peak detection, alignment and integration were optimised where possible.

In the final data analysis step, the metabolite $\times$ sample data matrix was analysed using

Figure V.2.: **Design of experiment:**

For details please refer to the text.

univariate and multivariate methods. Univariate methods include comparison of the intensity fold change to control, but also various correlation analyses. For multivariate analysis, unsupervised and supervised methods, such as principal component analysis (PCA) and partial least squares - discriminant analysis (PLS-DA) were used.

V.2. Methods

V.2.1. Sample generation and metabolite extraction

For an experiment of this scale, it is important to avoid batch effects, *i.e.* relating to the processing order of the samples, rather than strain specific effects. To prevent this, individual replicates and not strains were handled together for all the experimental procedures, including sample generation, extraction and acquisition. Three evenly spaced replicates of the wild type worms were processed within every replicate block.

V.2.1.1. Generation of longevity mutant screen samples

The long-lived *C. elegans* mutant strains used in this experiment are listed in table V.1.

Due to the high number of strains used, the growth and age synchronization protocols had to be adapted from the standard protocol outlined in B1. In order to minimise the risk of bacterial or fungal contamination, the agar plates did not contain any nutrient supplements, such as peptone. Instead, OP50 was grown as an overnight culture in 25 g/l Luria Broth medium, concentrated by centrifugation, and seeded directly onto the NGM plates (without peptone). One plate was seeded with 250 μ l of the equivalent of 25 ml LB culture (100 x). This was enough to keep worms fed *ad libitum* until adulthood. Before the final concentration and subsequent aliquoting step, all pre-concentrates were pooled to remove batch effects. Seeded agar plates were kept refrigerated until use within a week.

Due to the variation in developmental rates of the different strains, age synchronisation was modified in a way that allowed simultaneous growing of all strains and replicates.

Longevity mutants develop at different paces. Furthermore bleaching survival rates may be reduced drastically by certain mutations, particularly the ones affecting mitochondrial physiology¹. Thus, to obtain a high number of synchronised worms from all strains at the same time, plates set up at three consecutive days were pooled for synchronisation to obtain one biological replicate. At the time of synchronisation, 2 plates had been growing for 2.5 days, 3 plates for 3.5 days and 2 plates for 4.5 days. During synchronisation, the

¹Personal communication from M. Sumakovic, University of Cologne

pooled worms were bleached onto 8 plates.

To produce meaningful results, worms should be harvested at the same physiological time point. In most other experiments for this thesis, the developmental point chosen was either the L4 larval stage or early adulthood. Again, as longevity mutants develop at different paces and certain strains may even have a different morphological appearance during those stages, it was decided to determine the sampling time point differently. Worms were sampled at the first signs of the onset of egg laying, which corresponds to young adulthood.

Sampling was performed as outlined in B2 and frozen pellets (one per plate), stored at -80°C inside bead-beater tubes, pre-filled with 200 μl of 0.1 mm SiO_2 beads.

V.2.1.2. Generation of stable isotope annotation samples

All combinations of carbon, nitrogen and sulphur labelled *C. elegans* samples were prepared according to the protocol, developed in section IV.2.1.4 (None/C/N/S/CN/CNS). Cholesterol was added without ethanol, allowing full isotope enrichment. Three plates of worm biomass were grown for each isotope labelling combination, to allow the preparation of three extract aliquots. Worms were harvested in the same manner as the mutant screen samples.

V.2.1.3. Metabolite extraction

For the mutant screen, five out of the eight frozen pellets were extracted, as outlined in B2, using 80 % methanol and a bead-beater, yielding five replicates per strain. Three equal aliquots were prepared for each sample, to have the option to use different types of chromatography (at a later stage of this project). 10 % of the whole liquid extract volume were removed prior to aliquoting and dedicated to a pooled QC sample, which also was aliquoted.

The three replicates of stable isotope labelling samples were extracted the same way as the mutant screen samples, to guarantee identical coverage. All three replicates of the extract of a labelling regime were pooled and divided into three equal aliquots.

V.2.2. LC-MS acquisition

LC-MS acquisition was performed with the help of Matthew Lewis (Imperial College, London).

Samples were reconstituted with 100 μ l water. To solubilise the pellet, samples were sonicated for 10 min and vortexed. To remove insoluble debris, samples were centrifuged for 10 min at 10k g and 4°C. 90 μ l of the supernatant was transferred into flat bottom vials with silanised high recovery insert. Samples were kept inside a Waters Acquity sample manager at 4°C until injection by an Acquity UPLC.

Time	A[%]	B[%]	Flow[ml min ⁻¹]
0.00	99.0	1.0	0.600
0.10	99.0	1.0	0.600
14.50	3.0	97.0	0.600
16.50	0.0	100.0	1.00
17.50	0.0	100.0	1.00
17.60	99.0	1.0	1.00
17.70	99.0	1.0	0.600
20.00	99.0	1.0	0.600

Table V.2.: LC-MS RP gradient for final *C. elegans* metabolite annotation and longevity mutant screen

5 μ l of the sample was injected onto a Waters 2.1 $\times$ 100 mm HSS T3 column, filled with 1.8 μ m C18 linked silicon particles, that was kept at 45°C. At start, the flow rate was at 0.6 ml per minute and a 99 % A (A = 100% water with 0.1 % formic acid and 1 %B (B = 100% methanol with 0.1 % formic acid) solvent ratio. The solvent gradient is displayed in table V.2.

The LC eluent was introduced into a Waters Xevo G2 Q-ToF mass spectrometer. The eluent was evaporated and molecules ionised by an electrospray source heated to 120°C, with desolvation temperature at 600°C and a desolvation gas flow of 1000 l h⁻¹. The capillary voltage was set to 1.5 kV in positive ionisation mode. Sample and extraction cone voltages were 30 V and 4 V, respectively. 6 mass spectra per second were acquired over a range of 40 to 1200 Dalton and an effective scan time of 0.15 s. To obtain high accuracy mass measurements, the mass spectrometer was calibrated to sodium formate and a lock mass was introduced into the source throughout the run. The lock mass was 200 pg/ μ l Leucine-Enkephaline in a 1:1 mixture of methanol and water. It was infused at a rate of 10 μ l/min and recorded every 30 scans; averaging four lock mass scans each time. The whole system was controlled by MassLynx version 4.1.

After three blank injections, the column was conditioned, using the injection of 5 mutant screen QC samples. First, the isotope annotation study was run. The mutant screen samples were then run in replicate blocks, with QC samples every 12 samples.

Due to a college-wide, unplanned power outage, the acquisition was interrupted within the second replicate block of the mutant screen samples. As the power outage lead to a crash of the LC, but not the MS system, the injection series was resumed where it stopped, after re-conditioning the LC system with five QC samples. Thus, the acquisition was in effect a two-batch acquisition, requiring alignment of the two batches.

V.2.3. Processing and Analysis

After acquisition, all data were converted into the CDF format, using Waters MassLynx and peaks detected using xcms (version 1.33.14).

V.2.3.1. Processing of the isotope annotation samples

xcms parameter settings were kept liberal, to allow for the highest coverage possible. The "centWave" algorithm was used with an expected chromatographic peak width between 1 and 10 s and a signal to noise threshold of one. All other parameters were left at default values. The data reduction and annotation algorithm was used as described in the previous chapter. The retention time (rtE) and ppm accuracy ($ppmE$) thresholds, used for matching between the different labelling regimes were determined as described in IV.3.4.1. A $rtE = 1$ and $ppmE = 6$ were found to detect all biological peaks, visible in the BIP, but not detect any non-biological ones. $ppmE = 6$ was also used during the molecular formula calculation. For the atom-count plots, the M+1 adduct/formula pair was preferred for biological peaks, to which multiple adduct/formula pairs were assigned.

V.2.3.2. Database matching

The assigned adduct/formula pairs were matched to the compounds contained inside the PubChem, LMSD, KEGG and SMID databases.

The local PubChem copy from the simulations in chapter IV. was used. PubChem [56] contains separate entries for individual stereoisomers of a compound, such as L-alanine and

D-alanine, and an additional entry for the racemic form (DL-alanine). As stereoisomers cannot be separated by achiral chromatography, alanine, for example, should only produce one hit, when queried in a database. Thus, prior to matching, the local copy of the PubChem database had to be reduced to one entry for all stereoisomers, only. This was done comparing the canonical and isomeric SMILES (simplified molecular-input line-entry system) descriptors, of which the latter contains stereochemical information.

The Lipid Maps Structure Database (LMSD) [550] was downloaded on 01.05.2013 as CSV file and parsed into R using custom scripts. All lipid classes were considered, but restricted to CHNOPS compounds only.

A Kyoto Encyclopedia of Genes and Genomes (KEGG) [450] compound database version from 02.12.2010 was used. It was parsed into R using custom scripts. Only non-charged, non-salt CHNOPS compounds containing at least one carbon atom were considered.

The Small Molecule Identifier Database (SMID DB) was downloaded from smid-db.org on the 05.05.2013 and parsed into R.

The figure produced for the results chapter, which shows the matched adduct/formula pairs, was produced by first plotting all adduct/formula pairs as unmatched. The database matches were then overlaid in the following order: PubChem, KEGG, LMSD and SMID. Thus matches that were already matched by the previous database, are masked. However, this order represents the order of metabolite specificity, with KEGG and LMSD having a similar specificity, but different coverage.

V.2.3.3. Processing of the mutant screen samples

Before peak picking was performed, the *SN*, *prefilter*, *PPM* and *peakwidth* parameters of the xcms "centWave" algorithm were optimized. It was tested whether optimised parameters would reduce the number of unrelated, non-biological peaks and thus the total amount of data. The *SN* and *PPM* parameter can be found for each peak inside the peak list. The *peakwidth* may be calculated by subtracting the retention time where the ions of a peak were first (*rtmin*) and last (*rtmax*) detected. The *prefilter* parameter cannot be accessed directly. Instead, the EIC for each peak needs to be extracted. This can then be used to calculate for how many scans *k*, a certain peak was above the minimum inten-

sity *I*. ROC analysis was used to find whether predictive parameter settings exist, that would allow classification of peaks as biological. The ROC analysis was performed on the unlabelled sample from the isotope annotation workflow. Biological features were the gold standard for true ($n = 760$), and all other features, the gold standard for false classification ($n = 141409$). Based on this, a contingency table was calculated for each parameter and each of the biological parameter's settings ($n = 760$). From the contingency tables, the ROC curves were drawn. Before the actual ROC analysis, it was confirmed that the unlabelled sample from the isotope study and all samples from the longevity study could actually be compared. To this end, quantile - quantile plots were used, to ensure a similar parameter distribution of the peaks of those two sample(sets), when detected using the same peak detection parameters.

A loess model was used to correct the m/z domain between the two batches (pre/post power outage). The unlabelled sample of the isotope annotation study was also re-acquired during the second batch, so biological features could be matched between the two samples. All compounds that eluted within a five second window, sharing an identical carbon, nitrogen and sulphur count and an (unlabelled) mass within 60 ppm were considered identical between the two samples. The 60 ppm window was chosen as a worst case scenario, after manual comparison of several metabolites from both samples. A loess function with a smoothing factor of 0.7 was used to model the relationship between the mass difference of the same features from the two samples, with respect to their mass. The resulting calibration function was used to project the post-outage acquisitions onto the m/z scale of the pre-outage acquisitions.

In the next step, the peaks identified as biological, *i.e.* the target peaks, had to be found within the mutant dataset. A Monte Carlo (MC) simulation (1000 iterations) was carried out to determine the boundaries of the retention time and mass accuracy window used to find the targeted peaks within the samples of the longevity mutant dataset. The retention time window was varied ± 8 s around the targeted peaks' retention time; the mass accuracy window by ± 40 ppm. To improve simulation speed, the dataset was reduced to only contain data points within the maximum window size (8 s with 40 ppm) ($n = 216638$). The fraction of longevity mutant dataset peaks that was not yet matched by a single target peak and the fraction that was matched by one and only one target peak (unambiguous) was calculated for each iteration. Furthermore, the fraction of ambiguous matches, *i.e.* the

cases where a target peak was matched to several mutant dataset peaks, was determined. All simulations were performed on the Amazon EC2 cloud computing service in a parallelised fashion. For visualisation, a surface was fit to the data points using the R based "fields" package.

For retention time alignment a loess function was used as in the GC-MS alignment chapter (III.4) or as used by xcms. The median retention time profile was calculated from the unambiguously matched peaks. A loess model was constructed only using the retention time deviation to the median retention time of unambiguously matched peaks for each sample. The model was then used to correct all retention times of the whole peak list. For the first round of retention time alignment, a grouping window of ± 4 s and ± 10 ppm was used, as determined by MC simulation. A smoothing parameter ("span") of 0.65 was used, to correct for global trends. In the second round, ± 3 s and ± 10 ppm were used (as determined by another MC simulation, using the retention times corrected by the first round). The span was set to 0.3, to also allow the model to compensate for local drifts. The final grouping was performed using ± 2 s and ± 10 ppm (again determined by MC simulation).

The data matrix was back-filled where there were gaps in the data matrix. The MZ and RT boundaries were calculated for each sample. As all retention times were aligned to the median retention time, the median retention time of the start and end of a target peak was projected into the sample's original retention time, using the generated loess models. The MZ boundaries for the integration were the median of the target peak's MZ value ± 10 ppm. XCMS's *getPeaks* function was used to perform the integration.

After integration, the samples were normalized by probabilistic quotient normalisation, as described in section A3.2.2, to compensate for differences in sample concentration. The median intensity for each peak was used as a reference.

To guarantee reliable, high quality data, only metabolite candidates with coefficient of variation (CV) of less than 26 % in their QC samples throughout the run were retained for multivariate analyses.

V.2.3.4. Data Analysis

The initial principal component analysis (PCA) and partial least squares - discriminant analysis (PLS-DA) models, including all samples, were built in SIMCA P+ 13.0 (Umetrics, Umea, Sweden). The predictive variance, Q^2 , was computed by leave one out cross-validation.

To allow the calculation of p-values for the predictive quality and sample group separation, all subsequent PCA and PLS-DA models were built inside R [567], using the `prcomp` [567] and `plsreg1` [509] functions, respectively. Again, Q^2 was calculated by a leave one out approach. For all control vs mutant PLS-DA models, the Q^2 was associated with a p-value by using a permutation approach, as described in the literature [622, 89, 156]. For each "correct" model, a null hypothesis distribution was built, using 2000 models with a randomly permuted Y-vector. Another p-value was generated to gauge the quality of separation between control and mutant samples, using an approach recently described by Worley *et al.* [638]. This was done for both the PCA and PLS-DA scores, using a software tool ("PCA Utilities"), provided by Worley *et al.*. The parametric Pearson and nonparametric Spearman correlation between the p-values and Q^2 was also supported by permutation ($k = 10^4$).

Three multivariate PLS regression models with different Y variables were built:

Y_1 was the percent lifespan increase compared to control, as recorded in table V.1. Y_2 was the (non-parametric) ranked version of Y_1 . Y_3 was (non-parametric) categorical data, partitioning all samples into three groups, according to the magnitude of their lifespan extension (control (= 0%), moderate (< 50%), medium (< 100%), high(> 100%)). Importantly, Q^2 was calculated by a leave-one-group out method, rather than leaving out individual samples.

To assess how strains compare using univariate statistics, three hierarchical clustered (Euclidean distance) heat maps were drawn. First, a Welch's two sided t-test was used to calculate all p-values between a mutant strain and its control for each candidate metabolite. These p-values were then used for the first heat map, to see significant, but non-directional changes. No multiple testing correction was used as the p-values were used for visualisation and filtering only, rather than rigorous significance testing. For the second heat map, the fold changes between the median intensity of a metabolite candidate of each strain and its control were calculated and clustered as log 2 fold change. To improve the dynamic range

of the colour gradient, fold changes were limited to -8 and 8 fold. The third heat map is a combination of both heat maps. Again, the log 2 fold change heat map was plotted, but non-significant ($p \leq 0.05$) observations were set to zero fold.

It was also tested which metabolite candidates correlate with life expectancy (table V.1). Parametric Pearson and non-parametric Spearman rank correlation were used against the raw, unscaled intensity data. Overexpression strains were not included. A p-value was derived by permutation analysis for each correlation, on both a per sample and a per strain basis. For the per sample basis, the life expectancy vector was randomly permuted to assign each individual sample a life expectancy. For the strain based approach, the life expectancy was permuted on a per strain basis, *i.e.* all replicates of one strain always received the same life expectancy. In order to compare the correlated metabolite candidates with the metabolite candidates from the heat map, a custom coded heat map was plotted with the life expectancy correlation coefficient and p-value above the corresponding metabolite candidates.

As the most correlated metabolite candidates and the highest inter-strain fold change metabolite candidates were found to be the same, an attempt was made to assign them. Most of the metabolite candidates were late eluting. Thus, the remainder of the study was focused on lipids only. The lipids were grouped into classes of common oxygen, nitrogen and phosphorus count. By analysing in-source fragmentation (see appendix figure C4.4), it was found that nearly all observed lipids are glycerophosphocholines (phosphatidylcholines (PC)). By performing queries against the LipidMaps [550] and Chempidder [473] databases, using elemental compositions and also by the inspection of other in-source fragments and neutral losses (such as fatty acids), some of the lipid classes could be putatively assigned.

For a final analysis, it was decided to further examine the fatty acid carbon chain length and the degree of saturation, *i.e.* the number of double bonds, of two lipid classes: the diacylglycerophosphocholines (diAcGPC, positively correlated to lifespan), and the monoacylglycerophosphocholines (monoAcPC, negatively correlated to lifespan). The fatty acid chain length was calculated by subtracting the number of carbon atoms ($C_{AcPC} = 9$) contributed by the common glycerophosphocholine backbone. To calculate the number of double bonds, the carbon and hydrogen atom counts of the glycerophosphocholine backbone ($C_{AcPC} = 9$; $H_{AcPC} = 18$) were subtracted. Then, the number of carbon atoms from

the fatty acid's aldehyde groups and the additional proton from the fatty acid's methyl group were subtracted. For diacylglycerophosphocholines, this is $C_{diAcPC} = 2$; $H_{diAcPC} = 2$, for monoacylglycerophosphocholines, under consideration of the non-substituted glycerol-alcohol, it is $C_{monoAcPC} = 2$; $H_{monoAcPC} = 2$. Then the number of theoretical fully saturated hydrogens is calculated for the given carbon count, by multiplying the carbon count by two. The theoretical saturation hydrogen count can then be compared with the actual hydrogen count to determine the degree of actual saturation.

V.3. Results

The results section is divided into three main parts. The first two are methodological and the last one focusses on the biological results. In the first section (V.3.1), the workflow and tools developed in the last chapter were used to analyse a *C. elegans* extract, acquired by reversed phase, positive ionisation LC-MS. In a second section (V.3.3), the strategy of processing the longevity mutant dataset is presented. This section deals with optimising processing parameters, alignment and detection of biological peaks, producing a data matrix which can be used for statistical analysis. The final section (V.3.4), is concerned with the biological analysis of this data matrix. Univariate and multivariate approaches were used to find a common signature across strains associated with life expectancy. This inter-strain signature was based on concentration changes of particular lipid classes.

V.3.1. Data reduction and annotation of *C. elegans* features

Firstly, peaks of biological origin from a *C. elegans* extract were identified and annotated with a molecular formula, allowing subsequent database matching, using the method developed in the previous chapter.

Figure V.3 shows the base ion peak (BIP) chromatogram of the *C. elegans* extract, with peaks of biological origin highlighted, and an overlaid blank injection. Many high-intensity peaks are non-biological. This was confirmed by inspecting the corresponding raw data. As the blank demonstrates, this chemical "noise" does not originate from the chromatographic system, or the solvents, but from the sample itself. For worm culture and extraction, only high purity solvents and chemicals were used. However, plasticisers or polymers can be released from non-glass sample handling consumables, such as pipette tips, extraction and sample tubes [574, 216], particularly as organic solvents were used during extraction. It is not feasible to avoid all plasticware in an experiment such as this; furthermore, persistent organic pollutants (POPs) are ubiquitous.

The optimised matching parameters ($ppmE = 6$, $rtE = 1$) for the algorithm (see chapter IV.3.4.1) reduced the number of detected peaks by 99.3 %, leaving 931 biological features. 999 different molecular formula/adduct pairs were calculated for these, for 760 out of the 931 biological peaks. Thus, no adduct/formula pair could be generated for 171 of the

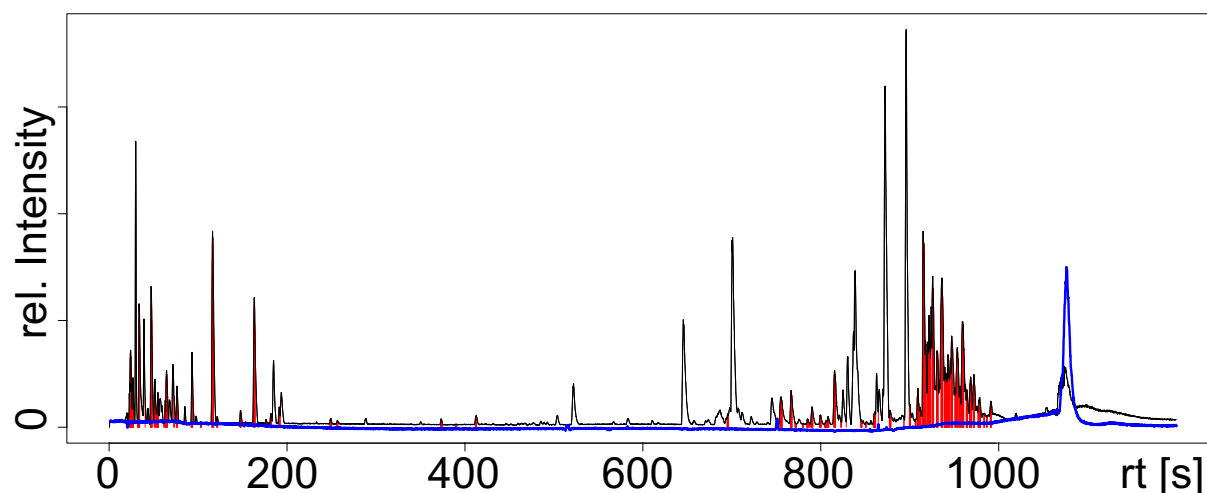


Figure V.3.: *C. elegans* BIP with detected metabolite candidates highlighted: The base ion peak (BIP) chromatogram of the *C. elegans* extract with metabolite candidates of biological origin highlighted in red. A blank (water) injection of the chromatographic system (scaled to match the same intensity axis) is overlaid in blue.

biological peaks. This may be due to a combination of reasons. It is possible that for some, the matching error window was too small. They may also have represented uncommon adducts, which were not considered by the formula generator. Even though unlikely, they could have exceeded the 99.99 % PubChem confidence intervals used by the formula generator (table C3.1 and table C3.2). There is also a possibility that some compounds contained elements other than CHNOPS. For example, selenocysteine is a known metabolite in *C. elegans*[200]. Another reason may also be a mismatch of labelling regimes, not immediately apparent in the check-tables.

Figure V.4 shows the different adducts assigned during molecular formula generation. As in the previous chapter, the selection of only a few very plausible adducts, out of 31 possible, shows the high specificity of this formula generation step. As expected, the M+H adduct was the commonest, followed by different alkali ones [44]. It was also checked whether metabolites had several different adducts, by finding identical molecular formulae (with different adducts), which eluted in a one second retention time window. 45 metabolite candidates were present with two different adducts (H/Na: 29; H/K: 10; Na/K: 6), and 27 with three different adducts (H/Na/K: 27; H/Na/2Na-H: 1). Even though Benton *et al.*[44] found the M+1 adduct to be the best pick, when deciding on an ion for quantification,

this cannot be assumed for each case. It is also possible that closely eluting isobaric compounds, such as lipids with different double bond positions of the fatty acids, which ionise as different adducts, were mistaken for different adducts of the same compound. Thus, the redundant "adducts" were not removed. Each adduct/metabolite pair will be quantified inside the longevity mutant dataset. However, during model analysis, their presence will have to be kept in mind.

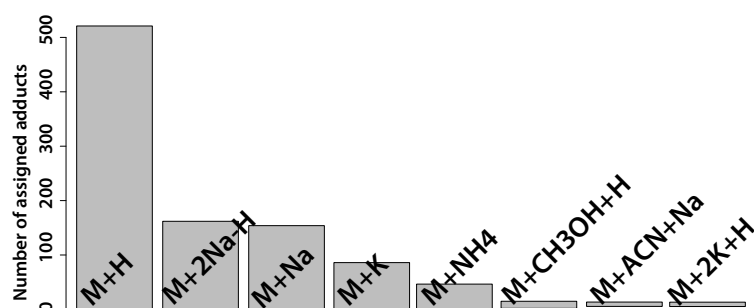


Figure V.4.: **Adducts detected in the *C. elegans* extract:**

Adducts and their count, assigned by the formula generator of the data reduction and annotation algorithm.

The average ppm error for formula calculation was -0.214 ppm, with a standard deviation of 2.30 ppm, and an average error of 1.84 ppm, if absolute ppm values were used. For none of the biological peaks (below 1 kDa) two different molecular formulae, but with the same adduct, *i.e.* $F_x + H$ and $F_y + H$, were calculated. This shows, that last chapter's simulation and conclusions seem to be correct: It is possible to calculate unambiguous molecular composition

at ToF accuracy, as soon as the carbon, nitrogen and sulphur count is known.

Figure V.5 shows the atom-counts of the molecular formulae that were assigned to the molecular peaks. All molecular formulae are highly plausible for the elution profile of reversed phase chromatography. Higher carbon-counts elute at later retention times. High molecular weight compounds eluting close to the void volume are polar, due to a high nitrogen count. Clusters of lipid classes, with clear presence of phospholipids, can be detected. The next section aims at querying the generated molecular formulae against different compound databases, to reveal their identity.

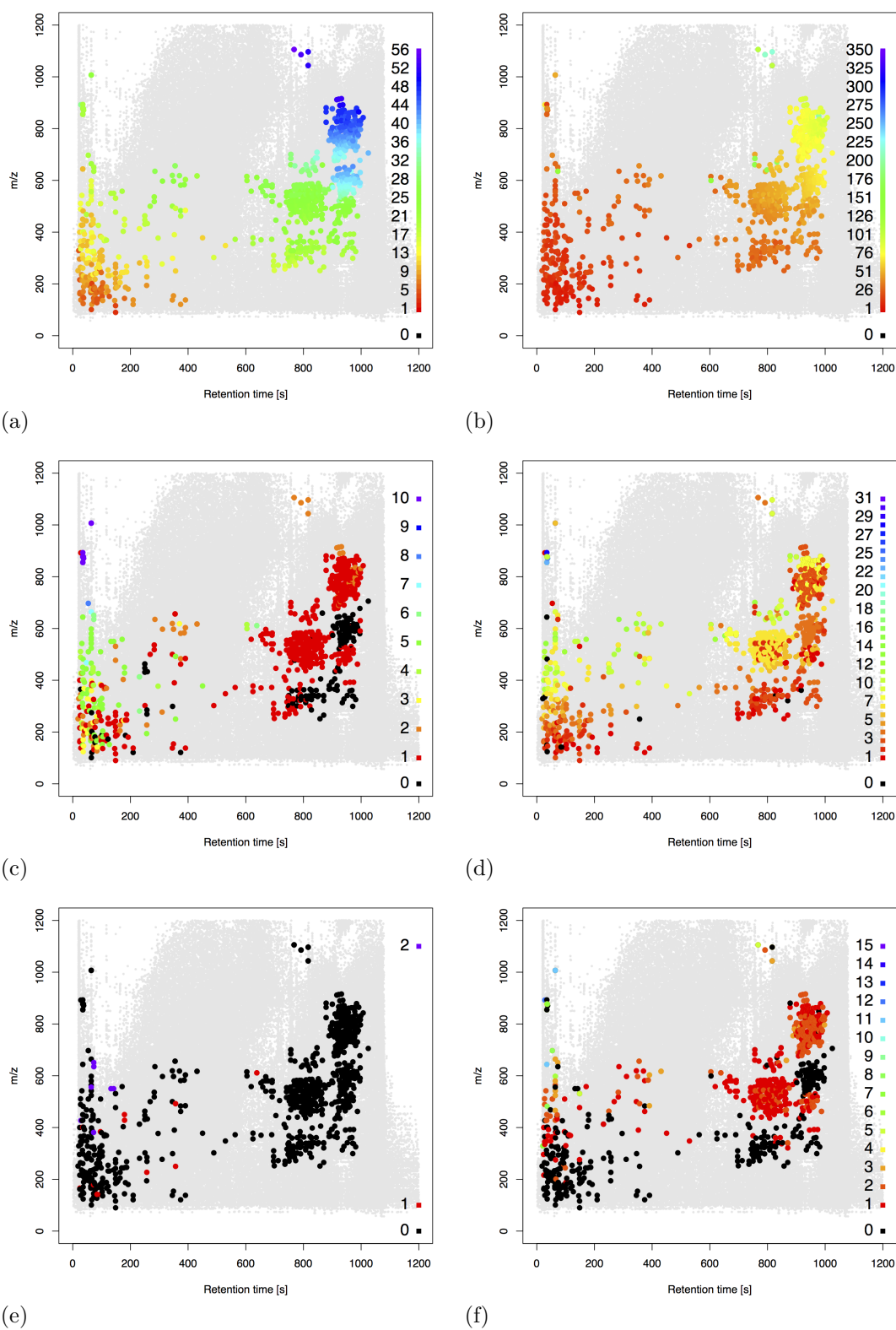


Figure V.5.: Molecular formulae of reversed phase ESI+:

RT vs m/z plots with atom-counts for C (V.5a), H (V.5b), N (V.5c), O (V.5d), P (V.5f), S (V.5e). All (non-reduced) detected peaks are in grey.

V.3.2. Database matching of *C. elegans* adduct/formula pairs

The isotope labelling strategy generates molecular formulae rather than assigned metabolites (i.e. metabolite candidates). Even though this often allows generalizations about the compound class, full metabolite assignment is preferred for biological interpretation. Several algorithms to calculate possible chemical structures from a molecular composition exist [163, 542], but chemical structure generation is a factorial problem and a single molecular composition can have billions of isomers [319]. Thus, an *ab initio* calculation approach is not useful. Therefore, the first step is to query a metabolite database.

In this section, this was done for the molecular formulae calculated above. Each composition may be assigned to several structures and therefore require experimental validation, according to the Metabolomics Standards Initiative's recommendations [551]. Therefore the focus is on what could be assigned and what is yet unknown, rather than the assignments themselves. Three categories of databases will be used:

PubChem [56] is likely to contain most of the scientifically reported compounds - thus it is used to assess whether it is likely that the compound is known.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [450] is one of the most complete and widely used metabolic pathway databases. Due to its wider coverage [14], it was preferred over MetaCyc [476]. As KEGG has a relatively poor coverage of lipid compounds, the Lipid Maps Structure Database (LMSD) [550] was additionally used to compensate for this. If a molecular formula can be matched to any of the two, it is likely to exist in a biological context, possibly with an associated metabolic pathway.

C. elegans also has some very organism specific metabolites. Ascarosides are sugar-based metabolites which act as important signalling molecules, particularly in the context of ageing and development [74, 536, 603, 276]. A large amount of different permutations of ascarosides with different functions exist [603]. They are deposited in the Small Molecule Identifier Database (SMID DB), which was therefore used to assign *C. elegans* specific molecules.

Figure V.6 and table V.3 summarise the database matching results. Even though the majority of formula/adduct pairs could be matched to at least one known (PubChem)

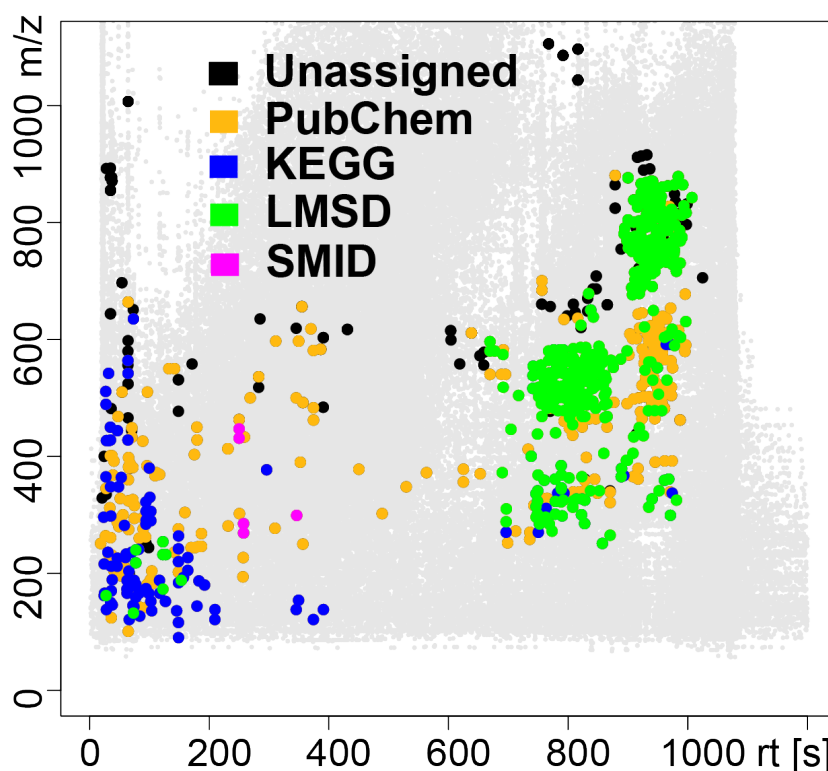


Figure V.6.: **Database matches for generated formula/adduct pairs:**

The *C. elegans* features, for which a molecular formula was generated, were queried against different databases. Details are given in the text.

compound, nearly a third (321) remained unannotated. Half (163) are from biological features, for which no formula/adduct pair has been matched yet. Thus, they are most likely undescribed compounds. As shown by the figure, most of them are of higher molecular weight, which may complicate elucidation. Lipid maps (LMSD) yielded the second most matches, unsurprisingly, as a major share of the detected features are lipids. Most of them are phospholipids, as confirmed by the LMSD metadata, but already known based on their molecular profile in figure V.5. On average, only KEGG and SMID delivered unambiguous matches. This may be mostly be due to the lower number of entries. For the SMID, it is also due to its high specificity, as it only matches *C. elegans* specific molecules. Four ascaroside molecules were matched, which corresponds well with expectations from previous studies [603], including retention times: they elute in a medium-polarity window where few other metabolites are present. The identification of these secondary metabolites confirms that the pipeline works well.

Database	Entries [n]	matched	"overmatch" [fold]
PubChem	16 594 399	678	18
LMSD	35 828	357	4
KEGG	13 592	145	1
SMID	164	5	1

Table V.3.: **Database matches for generated formula/adduct pairs:**

For each database, that the molecular formula/adduct pairs (n=999) were matched to, the total number of compound entries and number of matched compounds is given. Also the median of how many different database entries could be matched to a single formula/adduct pair ("overmatch").

The annotated peaks must then be extracted from the longevity mutant dataset. This is discussed in the next section.

V.3.3. Processing and matching of the mutant dataset

This section discusses the processing of the longevity mutant dataset, which ultimately leads to a metabolite candidate $\times$ sample matrix, which is then analysed in the next section.

V.3.3.1. Optimising the peak detection parameters

In order to reduce the total amount of data, it was investigated whether xcms peak detection parameters could be optimised (see section A3.1.3.1). Using the stable isotope reduction workflow, biological features were identified against a complex background. It is a plausible assumption, that biological features may have different analytical properties to other peaks. If this is the case, it may be possible to find specific peak-picking parameters which are selective towards biological features. The signal-to-noise ratio (SN), the minimum intensity threshold (*prefilter*), the ppm deviation between consecutive scans (PPM) and the peak width (*peakwidth*) can be adjusted for peak detection and may potentially help discern between biological and abiotic features. Receiver operating characteristics (ROC) analysis [164] was used. ROC analysis is used to select an optimal threshold for a binary classification algorithm (True, False). This is done by plotting the true positive rate

(TPR) against the false positive rate (FPR). The TPR, also referred to as sensitivity, is the probability of a classification as true, given that the questioned case is true. The FPR is the probability of classifying a case as true, even though it actually is false. For peak picking optimisation, true would be the classification "of biological origin", false "of other origin". The ROC plots are displayed in figure C4.2, in the appendix. The *prefilter*, the *PPM*, and the *SN* parameters were all of predictive value, when deciding between biological and not. They were all far away from the diagonal ($TPR = FPR$), which would indicate a random classification. The ROC curves are also coloured based on the percentage of expected data reduction a certain threshold could achieve. None of the parameters could allow for a reduction of more than 20 %, without at least risking a 20 % false positive rate. Thus, the benefits do not outweigh the risk of losing relevant information. Particularly as this experiment is semi-targeted, and so it was decided, to process the whole longevity mutant dataset with the liberal parameter settings, also used for the isotope annotation samples. This resulted in close to 16 million features in the whole dataset.

V.3.3.2. Alignment of the m/z domain

During acquisition, the injection series was interrupted by an unplanned power outage. This effectively led to a two batch sample acquisition. Even though the mass spectrometer did not reset, unlike the LC system, its absolute mass calibration was offset. As the lock-spray for the continuous mass correction assumed its proper function, the m/z domain of the second batch was still on a correct relative, but a wrong absolute scale. A loess (locally weighted scatterplot smoothing) regression function was used to model the difference between the absolute mass scales of the two batches. The regression function (shown in the appendix in figure C4.1), which is nearly linear between 200 and 1000 Da ($R^2 = 0.998$), was then used to post-acquisition correct the m/z values of all affected samples.

V.3.3.3. Dataset matching and retention time alignment

The next step was to find the metabolite candidates, *i.e.* the peaks of biological origin annotated with a molecular formula in section V.3.1, within the main dataset. This procedure corresponds to the grouping step in a peak alignment algorithm, such as provided by *xcms*. The use of *xcms* grouping function failed: the algorithm (*density*) did not find enough "well behaved peak groups" (see section A3.1.3.1), and the alternative

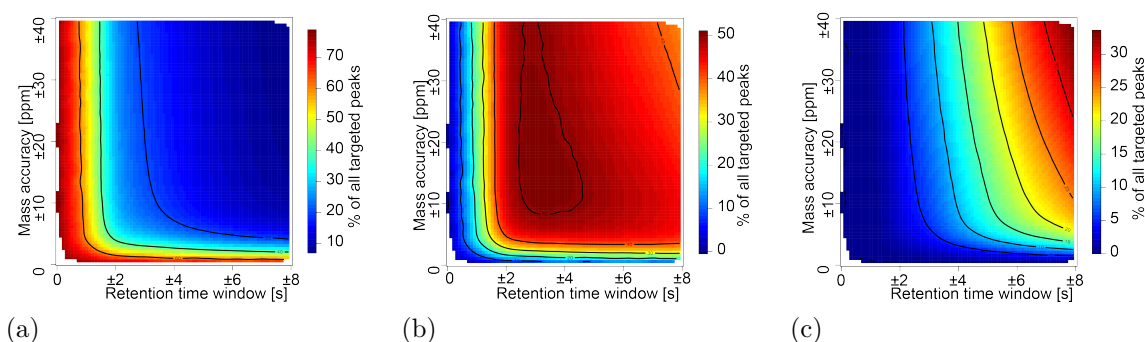


Figure V.7.: **Determining the matching window for the target peaks within the longevity mutant sample set I:**

The figure shows the simulation results, evaluating different retention time and mass accuracy window boundaries, to find the best thresholds to match the targeted metabolite candidates within all samples of the longevity mutant dataset ("grouping"). Figure V.7a displays the percentage of yet unmatched peaks. Figure V.7b shows the percentage of matches where a target peak matches one and only one dataset peak. Figure V.7c shows the percentage of peaks, which has already more than one peak that it could match to within the specified search window; *i.e.* its match is ambiguous.

algorithm (*nearest*) was aborted as only two of the 200 samples had been processed after two days. Thus, the off-the-shelf algorithms failed as a consequence of the sheer number of detected peaks. Therefore, custom functions were used to locate and group only the targeted metabolite candidates. In the first stage, the retention time and mass accuracy boundaries ("grouping window") were determined. This was done by performing a Monte Carlo simulation, evaluating different search window sizes, as done for the NMR database search in chapter III.2.3.3.

Figure V.7 displays the results of the simulation, used to determine the ideal retention time and mass accuracy boundaries for finding the target metabolite candidates. Figure V.7a shows the minimum boundaries required, by displaying the fraction of targeted peaks which do not yet have a match (from the sample set's peaks). Ideally, this number is very small. The minimum threshold is surprisingly clear cut for both dimensions. ± 3 ppm and ± 1.5 s would only leave 40 % of all (possibly matchable) sample set peaks unassigned. The thresholds at which peaks are found increase perpendicular to each other. This means they are independent and without major interaction. Figure V.7b shows the percentage of sample set peaks which has one and only one match from target peaks. This would be

the ideal case. Thus, a high percentage is desirable. Plot V.7c shows the percentage of ambiguous matches. These are those sample set peaks, which could be matched to more than one target peak. Thus, a low number can be considered as ideal. All three plots must be considered to select a threshold. Generally, it seems that the choice of the ppm threshold has a smaller influence on the result, than the retention time threshold. For the first round of grouping, before the retention time correction, it was decided to keep the settings more liberal. At 4 s and 10 ppm half of all peaks are unambiguously matched. Beyond that threshold, there is also little decrease of unmatched peaks and the percentage of ambiguous peaks is still rather low $< 10\%$.

Retention time correction for the grouped target peaks was also performed using custom written code. The algorithm is the same as the one used by xcms and in the GC-MS alignment chapter (III.4): a loess (locally weighted scatterplot smoothing) regression function to model each sample's retention time deviation from a median retention time across the whole run. Only unambiguous peaks were included. In figure V.8a, the proposed loess function for the retention time alignment to the median retention time can be seen for each sample. There are clear batch effects between the pre- and post-power outage samples and the unlabelled sample from the isotope study. Following the first round of RT correction, regrouping was performed (using a ± 10 ppm and ± 3 second window). A second round of retention time correction was conducted.

As shown by the MC simulation for the final grouping step, the retention time window can be reduced down to ± 2 s. Indeed, as shown by histograms, the distribution of retention time deviations from the median retention time profile has much improved from the first round of grouping (figure V.8b) to the one after the second round of retention time alignment (figure V.8d). Also, as figure V.8c shows, after the second round of retention time correction, barely any visible RT batch effects remain. Most of the individual peaks cluster close to the median retention time.

After the successful alignment of retention time profiles of the different samples, the full target peak by sample matrix was constructed. The gaps need to be filled, *i.e.* where a peak was not found in a particular sample. All gaps were filled using xcms functions, to make integration comparable to the one during xcms peak detection.

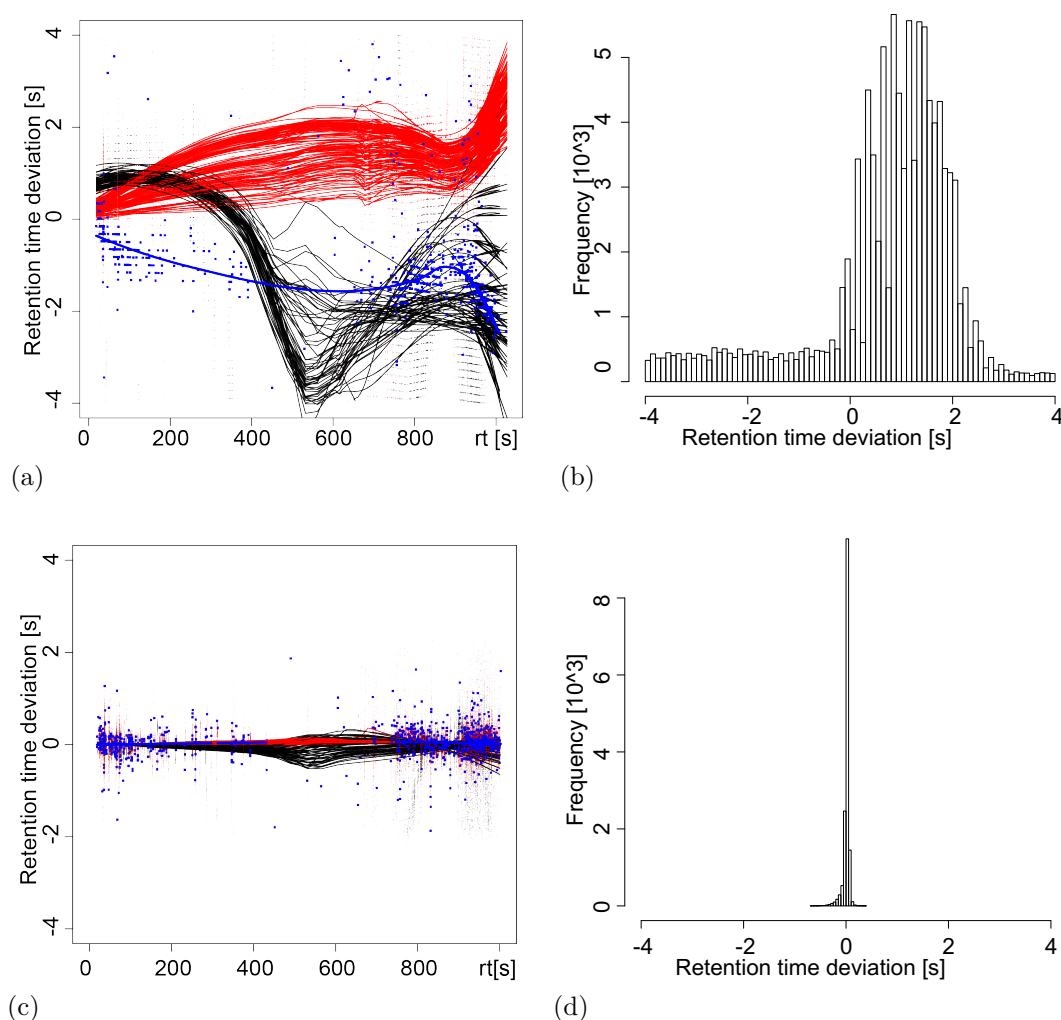


Figure V.8.: **Retention time alignment based on grouped, target peaks:**

Figure V.8a and V.8c show the loess regression models used for the first and (not actually applied) third round of retention time alignment, respectively. The models describe the deviation from the median retention time profile across the whole run length. The individual dots are the peaks, used to build the regression models. Peaks and models of the longevity mutant screen samples acquired before the power outage are coloured in black, the ones after the power outage in red. Blue is used for the unlabelled sample of the isotope annotation study, which was used to identify the target metabolite candidates. Histogram V.8b and V.8d show the distribution of deviations from the median retention time profiles before and after two rounds of retention time alignment, respectively.

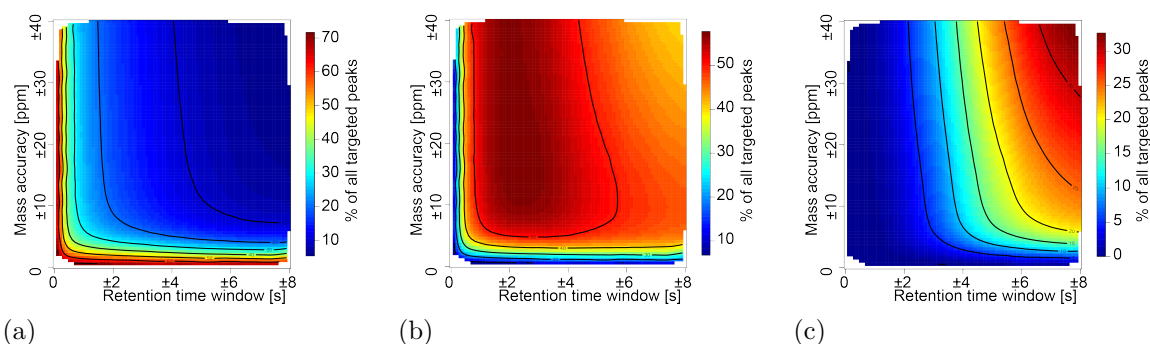


Figure V.9.: **Determining the matching window for the target peaks within the longevity mutant sample set II:**

The figure shows the simulation results for selecting the final retention time and mass accuracy grouping window (after the second round of retention time alignment). Figure V.7a displays the percentage of yet unmatched peaks. Figure V.7b shows the percentage of matches where a target peaks matches one and only one dataset peak. Figure V.7c shows which percentage of peaks has already more than one peak that it could match within the specified search window, *i.e.* its match is ambiguous.

V.3.3.4. Preparing the data matrix for analysis

Probabilistic quotient normalisation (PQN) was used to compensate for differences in sample concentration (see section A3.2.2). Figure V.10a and V.10b show box plots of the log₁₀ intensity distribution before and after PQN, respectively. A clear improvement in intensity homogeneity can be observed. Indeed, the median coefficient of variation (CV) of all QC sample metabolite candidates throughout the run improved from 36 % to 26 %.

As already discussed in chapter II.4.5, injection to injection intensity variation of the same sample during runtime cannot be totally avoided. Some metabolites ionise more reliably through the course of an acquisition than others. Thus, in large profiling studies it has become good practice to exclude those metabolites from (multivariate) analysis which cannot be measured reproducibly [141]. To do so, metabolite candidates which exceed a set CV threshold within their QC samples, are discarded [199]. For bioanalytical method validation, the Food and Drug Administration (FDA) recommends a cutoff at 15% and 20% if the analyte is close to the limit of quantification [165]. Here, the median QC CV (26 %) was selected as cut-off. Even though this value exceeds FDA recommendations, up to 30 % is considered acceptable for biomarker discovery studies [199, 296, 613].

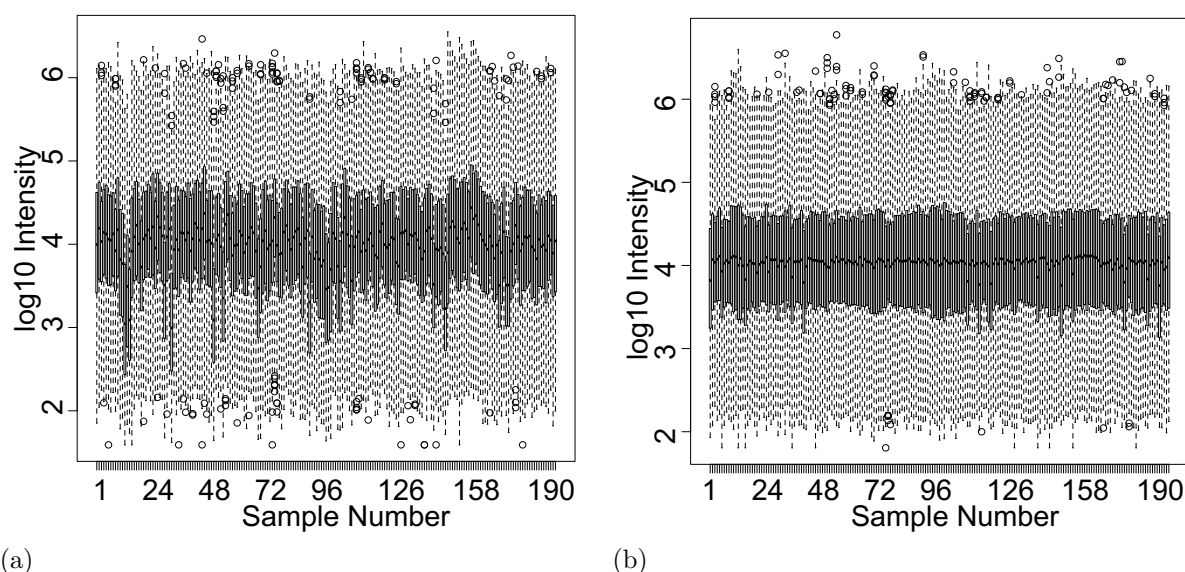


Figure V.10.: **Intensity distribution before and after intensity normalisation:** The box plots show the log10 intensity distribution for each sample before (figure V.10a) and after (figure V.10b) the probabilistic quotient normalisation. As reference, the median intensity for each metabolite candidate was used. Samples are not in runtime, but alphabetical order (with all replicates of a group next to each other).

Even though highly variable metabolite candidates have been removed, analysis may still be impacted by a runtime bias, such as a signal drift through the acquisition. To gauge the impact of unwanted, systematic bias, in relation to the samples, the clustering of the QC samples should be checked using unsupervised multivariate methods [511, 199, 296, 613]. Figure C4.3 shows a hierarchical clustered heat map of all samples. All QC samples cluster together, indicating a robust acquisition. No batch effects (pre/post power outage) are visible.

In LC-MS metabolite profiling, absolute signal intensity does not correspond directly to absolute concentration. Thus, absolute intensity information of individual peaks may bias a multivariate model. Therefore, it was decided to use unit variance scaling (see section A3.2.2) for all multivariate analyses. This gives all variables the same weight, regardless of their absolute intensity. It can be problematic with profiling data, as it may increase the contribution of (generally less intense) noise and artefact variables. However, as for the current study, all peaks are (mostly non-redundant) biological signals only, unit variance

scaling is appropriate.

V.3.4. Data analysis - the role of metabolism in *C. elegans* longevity

V.3.4.1. A single, all strain model, is not the best data analysis strategy

Could a single (multivariate) model be built for all long-lived strains to identify a common metabolic signature of longevity? To this end, an unsupervised (PCA) and a supervised model (PLS-DA) were built, combining all life-extending mutations, including their respective controls.

The resulting models can be seen in figure V.11. When inspecting the first two components of both models (figure V.11a and V.11b), a highly complex pattern of overlapping strains can be seen; in contrast to the desired clear separation between sample groups. One may even think that the two modelling techniques failed to reduce the dimensionality, *i.e.* complexity of the data, or that there is no metabolic difference between sample groups. This, however, is not the case. There are differences between strains, hidden inside the data matrix. When inspecting higher dimensional components (figure V.11c and V.11d), separation and clustering of individual strains is observed. The complexity of the models may simply be a consequence of the high number of individual sample groups (25). Both modelling techniques aim at reducing the dimensionality of the variables, by linearly combining them to "meta-variables" of maximum discriminatory power [579]. However, they do not summarise the dimensionality of the observations, for example by combining all replicates of one sample group. If the model complexity really is caused by the high number of sample groups, one would expect the number of latent variables required to sufficiently model the data, also to be high. For both the PCA and PLS-DA models, the cumulative explained variance (R^2) and the predictive value of this variance (Q^2) increases for more than the first 20 latent variables (figure V.11e and V.11f). After that, their predictive value stalls or even decreases, indicating overfitting [622, 61]. That the number of optimal components to describe the data is in a similar range to the number of sample groups, may not be a complete coincidence. Such a high number of latent variables, however, does not allow for straightforward model interpretation. Thus, this data set illustrates the limitations of those two powerful techniques. It also requires a different strategy for data analysis. It may be appropriate to analyse smaller models, which describe parts of the problem and can

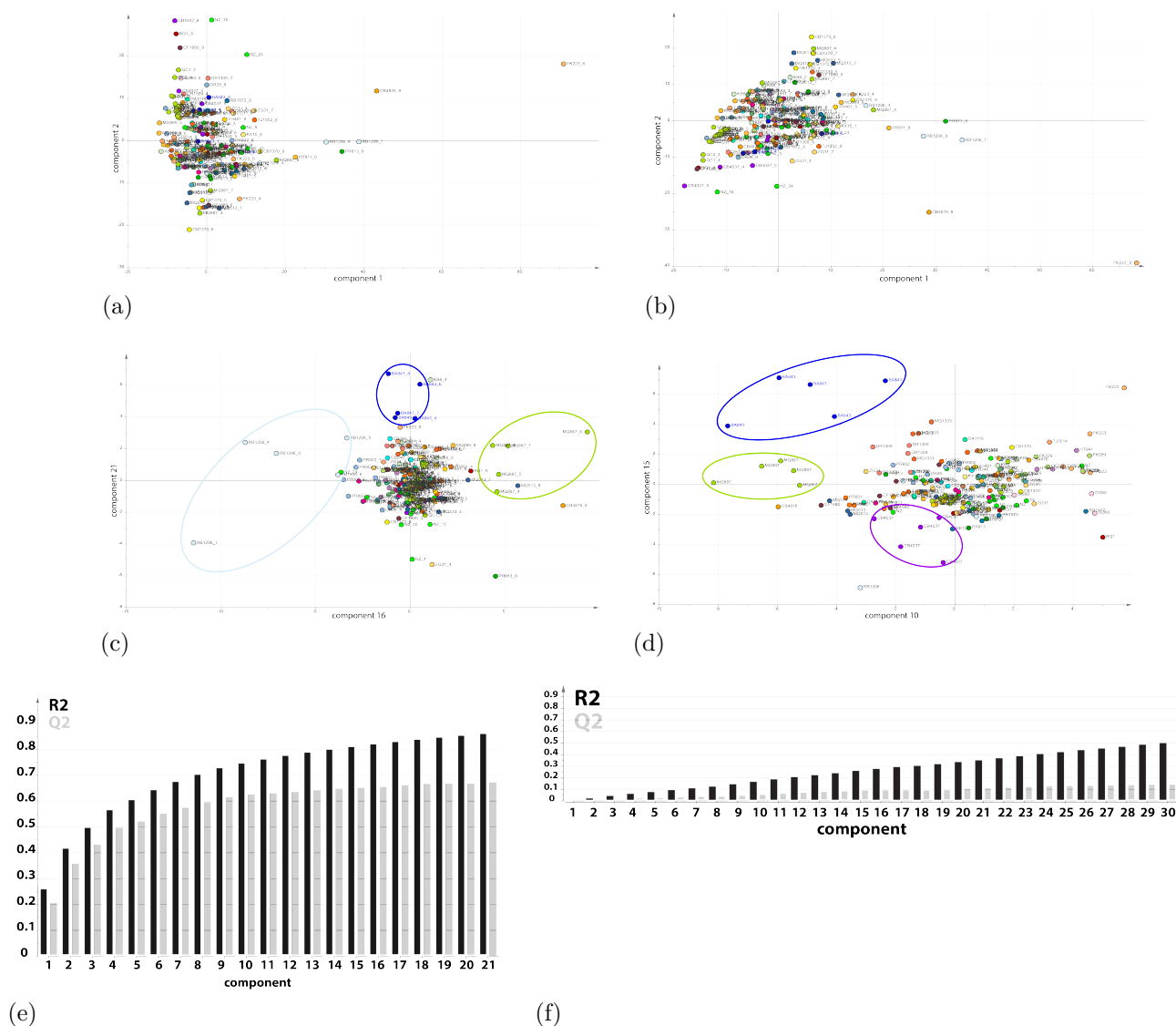


Figure V.11.: **PCA and PLS-DA models including all samples:**

Figure V.11a, V.11c and V.11e are from the PCA model. Figure V.11b, V.11d and V.11f are from the PLS-DA model. In figure V.11a and V.11b the first two components of each model are shown. V.11c and V.11d display higher components, in which several strains form distinguished clusters. In figure V.11e and V.11f the cumulative explained variance (R^2) and cross validated (predictive) variance (Q^2) can be seen for each of the fitted components (latent variables) of the two models.

then be summarised to gain a holistic picture, rather than creating a model of all strains.

V.3.4.2. Most mutant's metabolic phenotype differs from the control

The hypothesis is that individual mutants have significantly different metabolite concentrations, compared to the control. To test this, two-group models comparing all replicates ($n = 5$) of a knockout mutant strain to wild type ($n = 15$) were evaluated. For the Sir2.1 and cJNK overexpression mutants, the appropriate control ($n = 5$), different from wild type, was used. Again, PCA and PLS-DA were used as unsupervised and supervised models. PCA separates samples based on maximum variance. Thus, a separation between control and long-lived strain (within the first two principal components) indicates that there is more in-between than within group variation. PLS-DA can answer whether sample group membership can be predicted, based on the metabolite profiles. The predictive value of a model is expressed by Q^2 , a cross validated version of the R^2 score. The closer it is to one, the more predictive the model. Thus, a positive Q^2 value only tells how much of a model's variance is predictive. It does not make any statement about the likelihood of a possible mis-classification event. Therefore, the Q^2 value of each model was converted to a p-value, using a permutation approach ($n = 2000$); a strategy described, and recommended in the literature [622, 89, 156]. Furthermore, a p-value based method, recently described by Worley *et al.* [638], was used to determine, whether the separation of two sample groups is statistically significant.

Table V.4 summarises all multivariate models, comparing the individual strains of this study to control. Overexpression strains LG394 and HT941 were compared to the genetic background matching control strain LG398, whereas all others used wild type as control. This was necessary as N2 and L398 have a very distinct baseline metabolome, as their PCA separation indicates ($p \leq 6.21 \times 10^{-08}$). As expected, PLS-DA, as a supervised method, had greater separation of the sample groups. The predictive power of each model was tested, using a permutation derived p-value. All models did pass with a significance level of at least $p \leq 0.05$, and frequently much better. Thus, as hypothesised in V.3.4.1, the dataset contains a wealth of information, which simply cannot be captured inside a single model and expressed in two/three dimensional projections. Individual models, however, may reveal this information.

gene (allele)	PCA p	PLSDA p	PLSDA Q^2_{cum}	PLSDA Q^2 p	LSE [%]
<i>spe-26</i> (it118)	6.1E-01	2.4E-13	0.88	$\leq 5.0E-04$	45
<i>akt-1</i> (mg306)	3.7E-01	7.3E-08	0.52	5.0E-03	48
<i>daf-2</i> (e1370)	2.0E-07	1.8E-12	0.91	$\leq 5.0E-04$	100
<i>unc-64</i> (e246)	5.5E-01	1.4E-09	0.76	$\leq 5.0E-04$	35
<i>glp-1</i> (e2141)	3.3E-03	2.0E-10	0.76	$\leq 5.0E-04$	75
<i>clk-1</i> (e2519)	1.1E-03	2.4E-10	0.83	$\leq 5.0E-04$	40
<i>eat-2</i> (ad1116)	1.3E-01	2.2E-07	0.64	$\leq 5.0E-04$	40
<i>egl-4</i> (ks61)	5.6E-01	2.5E-09	0.82	$\leq 5.0E-04$	25
<i>jnk-1</i> (+)	7.6E-01	3.9E-05	0.37	4.00E-02	25
<i>rle-1</i> (tm2447)	2.0E-04	2.0E-11	0.87	$\leq 5.0E-04$	25
<i>ife-2</i> (ok306)	2.6E-02	5.6E-10	0.82	$\leq 5.0E-04$	25
<i>sir-2.1</i> (+)	3.6E-01	1.3E-04	0.58	1.8E-02	15
<i>nuo-6</i> (qm200)	3.8E-01	7.4E-10	0.83	$\leq 5.0E-04$	70
<i>daf-2</i> (e1370); <i>clk-1</i> (e2519)	1.1E-05	2.8E-13	0.92	$\leq 5.0E-04$	400
<i>isp-1</i> (qm150)	1.9E-02	1.9E-13	0.89	$\leq 5.0E-04$	65
<i>isp-1</i> (qm150); <i>ctb-1</i> (qm189)	9.4E-03	3.4E-14	0.88	$\leq 5.0E-04$	60
<i>tax-4</i> (p678)	7.4E-03	5.2E-10	0.84	$\leq 5.0E-04$	90
<i>che-3</i> (p801)	1.7E-01	3.4E-07	0.62	$\leq 5.0E-04$	100
<i>osm-3</i> (p802)	4.4E-03	1.2E-11	0.90	$\leq 5.0E-04$	65
<i>osm-6</i> (p811)	9.5E-01	7.1E-09	0.55	$\leq 5.0E-04$	60
<i>osm-5</i> (p813)	4.2E-02	1.3E-12	0.79	$\leq 5.0E-04$	120
<i>sod-2</i> (ok1030)	4.2E-03	5.5E-10	0.81	$\leq 5.0E-04$	50
<i>rsk-1</i> (sv31)	2.4E-03	1.8E-10	0.77	$\leq 5.0E-04$	25
<i>age-1</i> (hx546)	7.0E-01	3.5E-08	0.26	4.3E-02	100
<i>ifta-2</i> (tm1724)	4.4E-01	2.5E-07	0.55	1.5E-03	30
<i>hif-1</i> (ia4)	4.3E-01	4.5E-11	0.78	$\leq 5.0E-04$	25

Table V.4.: **PCA and PLS-DA models of mutant versus control:**

This table summarises the performance of all PCA and PLS-DA models. For each model two components were fitted and each strain was compared to a control strain (such as wild type). P-values describing the model's confidence for the separation [638] of the two strains are given for PCA and PLS-DA. For the PLS-DA model, the predictive variance (Q^2) and associated p-value is also listed. Furthermore, the lifespan extension for each strain is given. For the convenience of interpretation, cells have been coloured. Significant p-values are highlighted in red with $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.001$. Q^2 is on a continuous blue colour scale from 0 to 1. The lifespan extension is on a continuous green colour scale, ranging from 0% to 100%.

		PCA p	PLSDA p	PLSDA Q^2_{cum}	PLSDA $p(Q^2)$
parametric	R	-0.21	-0.14	0.18	-0.03
parametric	p-value	0.12	0.10	0.14	0.44
nonparametric	ρ	-0.17	-0.26	0.30	-0.11
nonparametric	p-value	0.18	0.08	0.05	0.29

Table V.5.: **Correlating the model quality and lifespan extension:**

Parametric (Pearson) and nonparametric (Spearman) correlation of the model parameters and % lifespan extension, of the values from table V.4.

It was also tested whether any of the parameters describing a model's performance, such as the Q^2 or p-values, correlated with the % lifespan extension. The results are summarised in table V.5. On average, the nonparametric correlations were better than the parametric correlation. Still, nearly all correlations were very weak and non-significant, as determined by permutation tests ($n = 10k$). The nonparametric correlation of the PLS-DA Q^2_{cum} , however, appears to be significant at $p \leq 0.05$ and positively correlated, with a Spearman's rank correlation coefficient of $\rho = 0.30$. This would mean that a more predictive metabolic phenotype at youth is connected to a longer lifespan extension. Even though the weak correlation precludes any strong conclusions, the potential correlation of lifespan extension to the metabolic phenotypes deserves further attention.

V.3.4.3. The multivariate link between life expectancy and metabolism

The major goals of this chapter is to evaluate whether there is a global signature or even a single biomarker for life expectancy.

The weak, but significant nonparametric correlation analysis between the predictive quality (Q^2_{cum}) of the individual PLS-DA models and life expectancy should be further investigated. To this end, a PLS regression against the lifespan extension for each mutant was performed. The overexpression mutants were not included, as they would require a separate control, which was shown to be metabolically distinct from the regular control. Three regression models were built. The first one aims to model the absolute lifespan extension (figure V.12a), expressed in percent from wild type. The second model is non-parametric, as it regresses to the ranked lifespan extension (figure V.12b). For the third model, the strain's life expectancy was divided into an ordinal variable with four ranks: control (= 0%), moderate (< 50%), medium (< 100%), high(> 100%).

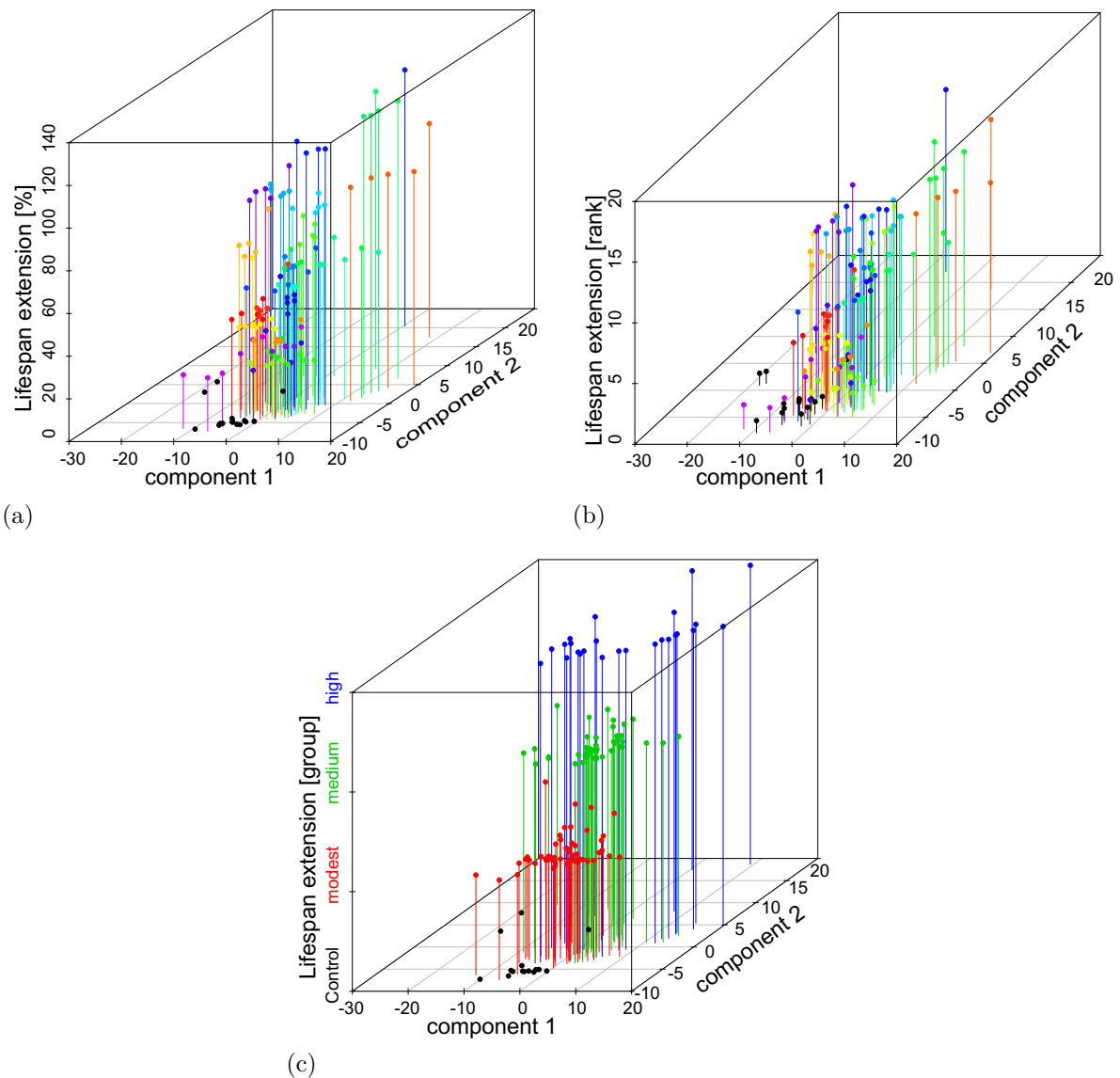


Figure V.12.: **PLS regression models of life expectancy:**

PLS regression of absolute (figure V.12a) and ranked (figure V.12b) lifespan extension. Each strain is marked in a different colour; black is wild type. In figure V.12c, the life expectancy is partitioned into three groups (control (= 0%), moderate (< 50%), medium (< 100%), high(> 100%)), to qualitatively model the magnitude of life span extension.

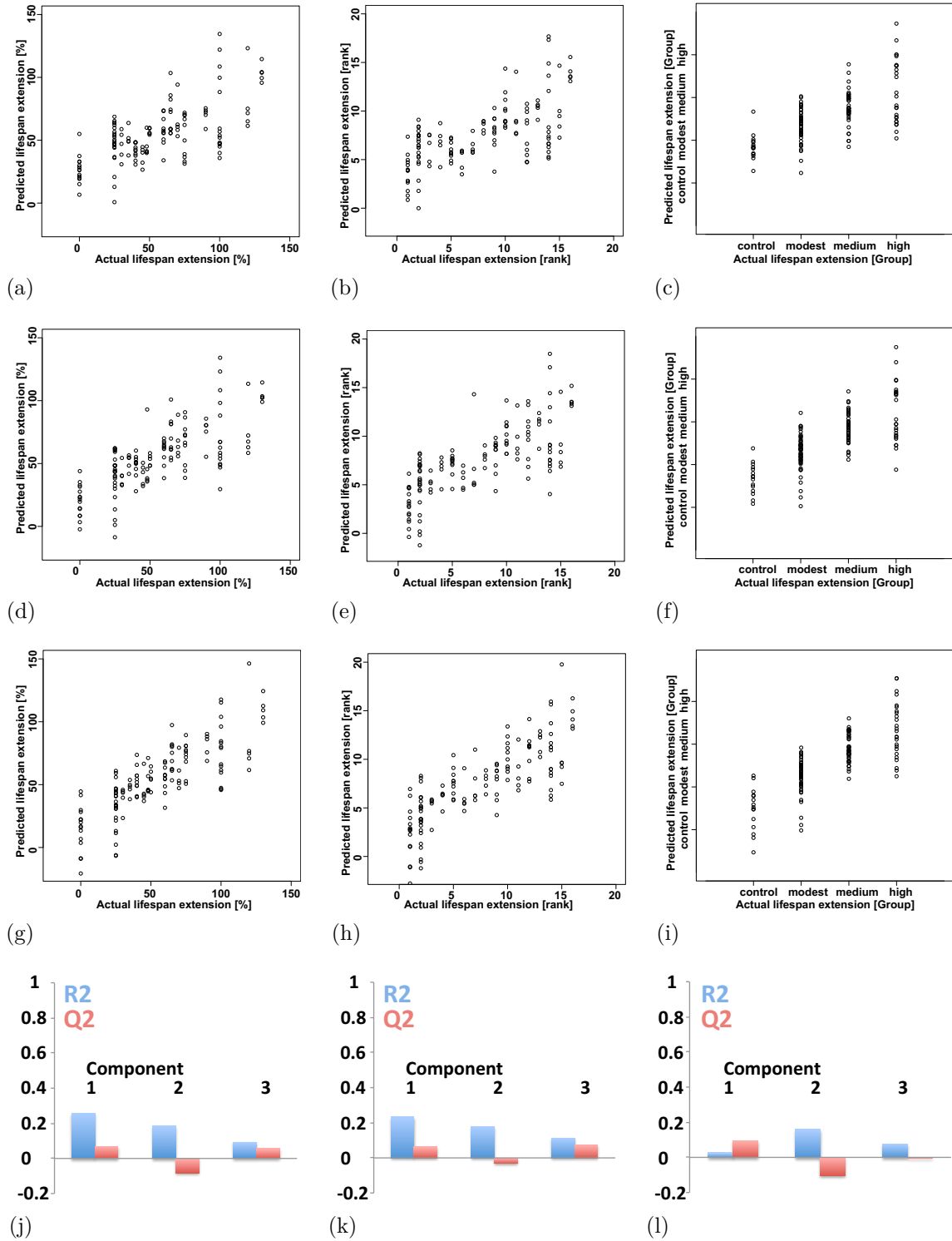


Figure V.13.: Predicted versus actual lifespan extension:

The predicted (y-axis), versus the actual lifespan (x-axis) is plotted. The first column is from the absolute lifespan extension model, the second from the ranked and the third from the group one. Row one is component one, row two, component two and row three is component three. The last row of plots shows the R^2 and Q^2 values for each of the model's components.

All three models, in particular the two rank models, show a clear trend (along their diagonal). As expected, not all of the 25 included long-lived strains follow this trend; also some (2-3) replicates of the wild type control strain show outlying behaviour. To further investigate this trend, the predicted, versus the actual life expectancy was plotted for each of the three models for component 1 to 3 (figure V.13). Indeed, the life expectancy is predicted reasonably well. As can be seen in figures V.13j, V.13k and V.13l, all models explain $\geq 40\%$ of the variance (R_{cum}^2) across the first two components. However, none of the models passed the cross validation step, with their Q^2 close to zero or even being negative. This means the lifespan extension can be modelled, but the models are not robust and should therefore not be consulted for biological interpretation [622]. As all data for lifespan extension was taken from literature and therefore not acquired under standardised conditions, it is hardly surprising that a model using absolute numbers may not be robust. Thus, the two PLS against rank models were built, in which the leverage of some extremely long-lived mutants, such as the *daf-2; clk-1* double mutant, is reduced. These models were expected to be more robust, as they should help compensate for problems arising from the comparison of absolute lifespan extension data across different labs, as such data tend to be variable [108].

Thus, it may be concluded, that no robust, multivariate metabolic predictor of longevity can be found using the employed method. It is very likely, however, that removing sample groups from this model will increase its predictive abilities and lead to such a signature for certain subgroups, as already demonstrated for insulin like signalling [183, 392] and mitochondrial [77, 75] mutants. The global PLS models are trying to model too many different factors. The next obvious step is to consider univariate methods. As individual biomarkers may even be preferred over multivariate signatures for biological interpretation, it was decided to find lifespan-correlated metabolite candidates.

V.3.4.4. Inter-strain comparison shows a distinct, systematic metabolite signature

The most basic univariate method may be a simple significance test, such as a t-test. These are often used as an initial filter for omic data, to reduce the number of variables to only statistically relevant ones [582, 435, 120]. During a first, coarse filtering step, type one errors, arising from multiple testing may be accepted in order to avoid type two errors. This

is perfectly justifiable, if any hits will be subject to rigorous statistical follow up testing.

Figure V.14a shows a strain versus variable (metabolite) heat map. The colour corresponds to the (log 10) p-value of a t-test between the mutant and its corresponding control, for any given variable. Clearly, as already gauged from the individual PCA and PLS-DA models, certain strains are more distinct from wild type than others. Any clusters in this heat map are non-directional, *i.e.* they are grouped by affected metabolite candidates, rather than by clusters of up or down regulation. Figure V.14b shows the (log 2) fold change of each mutant's variables versus control. Some systematic metabolite patterns across most strains can be seen. Several metabolite candidates are increased or decreased by over 8 fold. However, not all strains seem to be subject to these systematic concentration changes to the same degree. In fact, when only fold changes which are significant (t-test $p < 0.05$) are considered, some of the strains appear to be fairly unaffected. This can be seen in figure V.14c, which is a combination of figure V.14a and V.14b and assigns all non-significant (t-test $p > 0.05$) comparisons a zero fold change.

Overall, there appears to be no pattern correlated to DAF-16 dependence, nor clear functional clustering of genes acting by the same mechanism/pathway of lifespan extension. However, some do cluster, such as the mitochondrial (MIT) mutants. Surprisingly, the insulin like signalling (ILS) mutants do not. The ILS; MIT (*daf-2; clk-1*) double mutant and the *daf-2* mutant itself show the most consistent and distinct signature. The *clk-1* signature itself is far away, indicating a dominating contribution of the ILS pathway to this extremely long-lived double mutant. Generally, mitochondrial and ILS mutants appear to be very similar in their metabolic response, which is intriguing, given their assumed independence. The lifespan of the SOD mutant has also been challenged. However, it shows a distinct metabolic response, similar to the one of many other lifespan extending mutations, such as ILS and MIT. Also the TOR-dependent translation mutants, particularly the *rsks-1* mutant, have distinct patterns. It shares parts of the common signature, but it also has unique features, which is in agreement with its independence of most other known life-extending mechanisms. Many of the less well-characterised sensory mutants produced a signature which aligns with the overall pattern, confirming that they eventually feed into classic downstream longevity mechanisms. Knockout of *rle-1*, which degrades the key longevity regulator DAF-16, leads to a very similar pattern as knockout of the DAF-16 suppressor *daf-2* itself. Still, it is not clear why other members of the ILS cascade (such

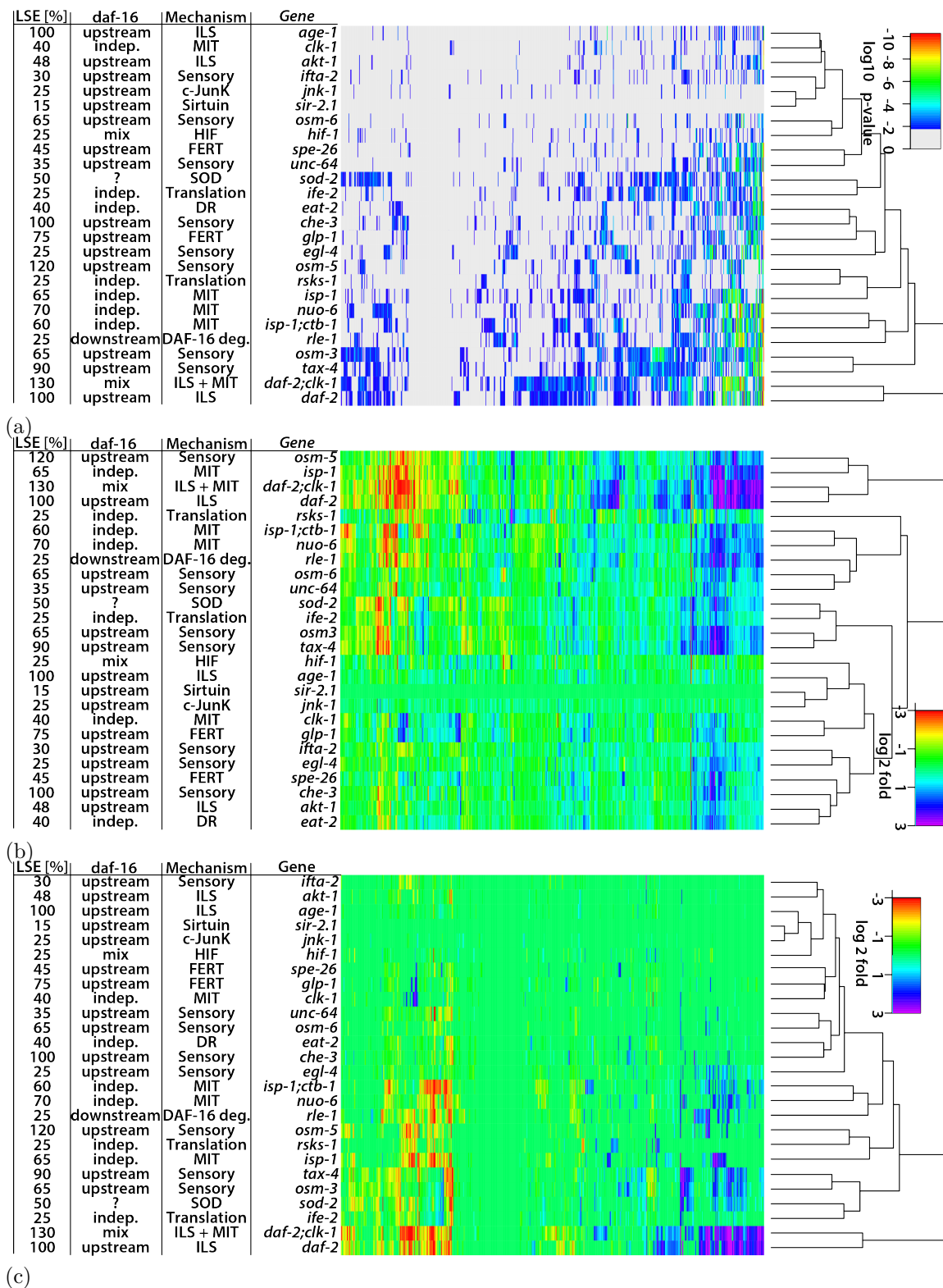


Figure V.14.: Clustered heat maps of (significant) fold changes:

Figure V.14a: All metabolite candidates significantly different from control (t-test $p < 0.05$); figure V.14b: Fold changes of all metabolite candidates (mutant vs control); figure V.14c: Combination of both figures: Only significant (t-test $p < 0.05$) fold changes are considered. For details, please refer to the text.

as *age-1* or *akt-1*) failed to reproduce the *daf-2* signature. However, some other mutants also had little to no significant metabolite changes, e.g. the overexpression mutants *jnk-1* and *sir-2.1*. Also the fertility (FERT) and dietary restriction (DR) mutation did not show a strong phenotype, but still had some metabolites which followed the overall pattern.

Overall, however, there certainly appears to be structure in the dataset, with commonly up and down regulated metabolite candidates. But what is its cause? One property all mutants have in common is modulation of life expectancy. Thus, in the next step, it was tested whether some metabolite candidates had a direct correlation with mutant life expectancy.

V.3.4.5. The inter-strain metabolite signature is correlated to life expectancy

In the above analysis, a common pattern of co-regulated metabolite candidates across many of the strains was detected. It was hypothesised that this could represent a global signature, correlated to life expectancy. Hence, all individual metabolite candidates were correlated to life expectancy. Figure V.15 shows the results. (The overexpression mutants were excluded.)

Both, parametric Pearson correlation (figure V.15a and V.15b) and the more conservative non-parametric Spearman rank (figure V.15c and V.15d) correlation were used. All correlations were also validated by assigning permutation ($k = 10^4$) derived p-values to each correlation. Two permutation approaches were used. Individual sample (figure V.15a and V.15c) and group-wise (figure V.15b and V.15d) permutation. For the individual sample permutation, the lifespan extension vector was permuted randomly. When group labels were permuted, all replicates of one strain received the same lifespan extension. Thus, the most liberal and conservative scenarios are displayed in figure V.15a and V.15d, respectively. Many metabolite candidates show a clear and statistical significant correlation with life expectancy. Both positive (direct) and negative (indirect) correlations can be seen.

Figure V.16 shows the heat map, but also gives the correlation coefficient and p-value of the correlation for each metabolite. It also indicates whether the metabolite is a lipid. It can be seen that metabolite candidates subject to the global fold change pattern also have significant correlation to lifespan. This confirms the hypothesis that the inter-strain pattern, which dominates the dataset, describes the link between metabolite concentra-

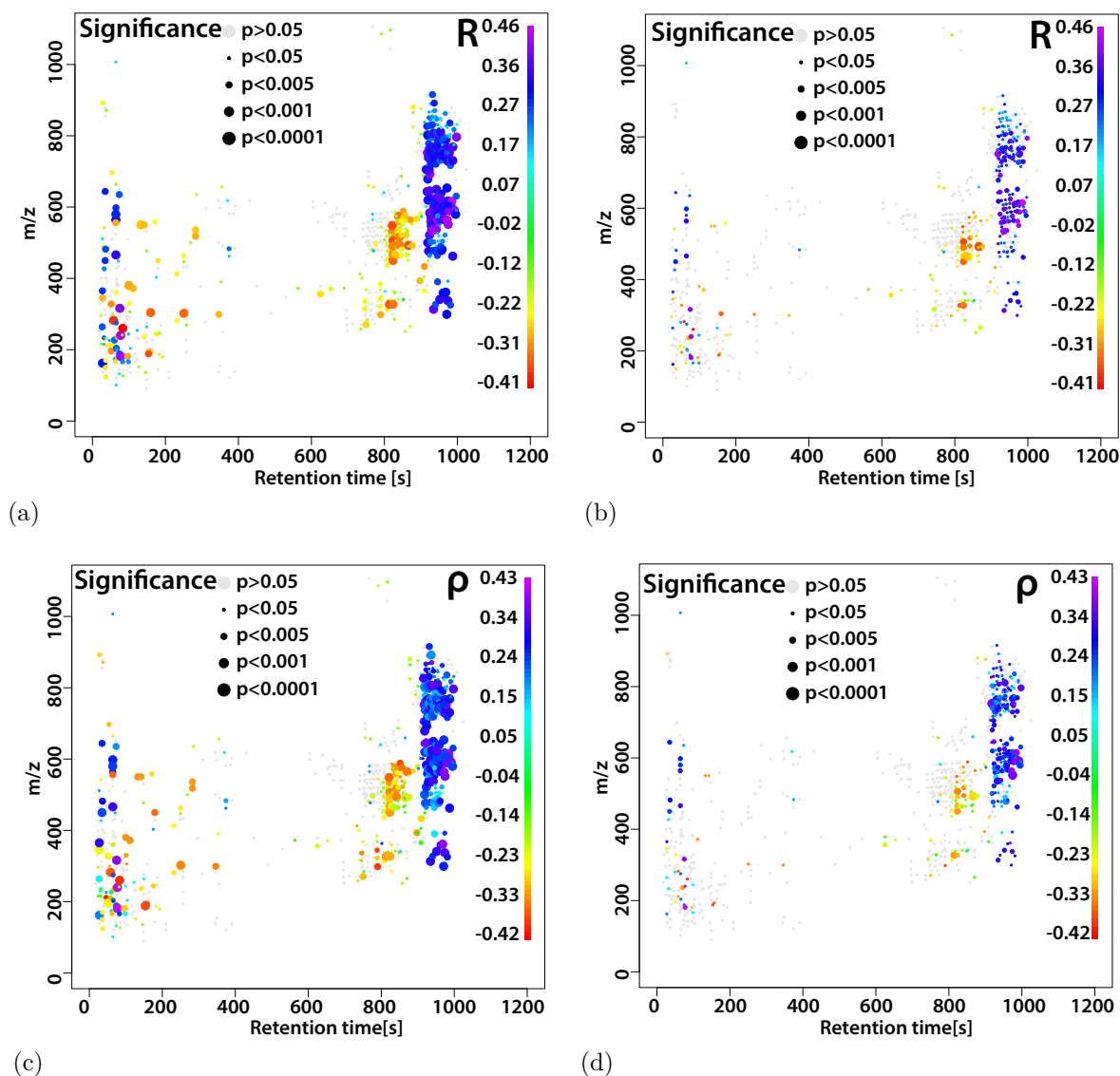


Figure V.15.: **Univariate intensity - lifespan correlation for each metabolite:**

The correlation coefficient between intensity and lifespan extension was calculated for each metabolite individually. In figure V.15a and V.15b the Pearson correlation coefficient (R) was used. The nonparametric Spearman rank correlation coefficient (ρ) was used in figure V.15c and V.15d. The colour scale corresponds to the correlation coefficient's magnitude and direction. The size of each scatter point increases with the correlation's significance, expressed by its p-value. To calculate the p-value, the individual label (V.15a and V.15c) and group (V.15b and V.15d) permutation approach, described in the last section, were used. All grey data points did not reach a significant p-value ($p \leq 0.05$).

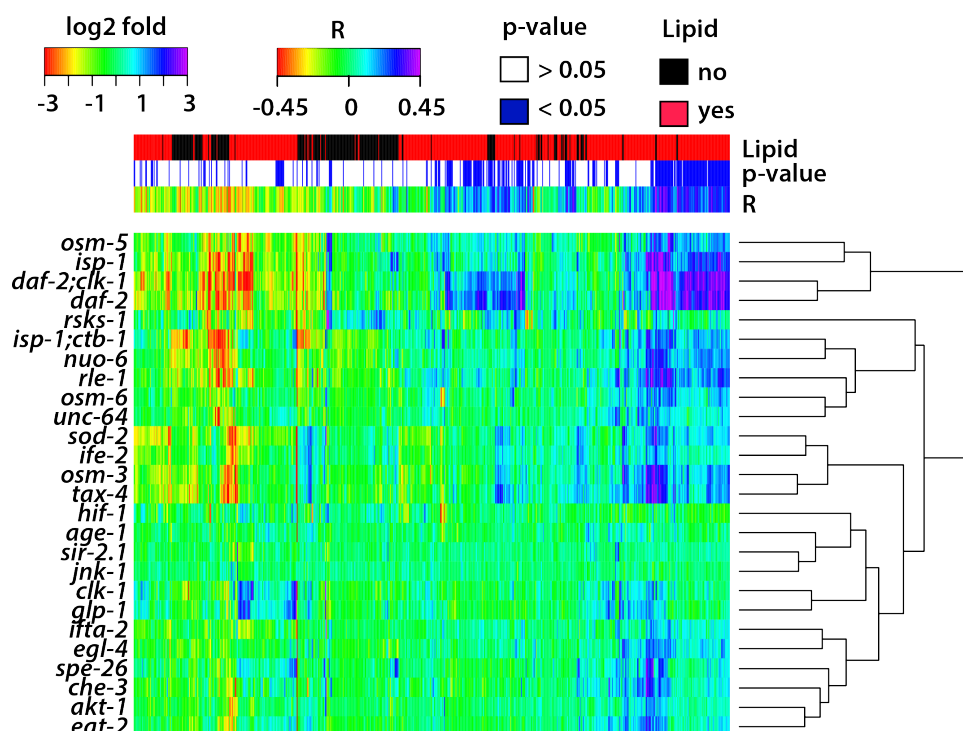


Figure V.16.: **Intensity fold change and life expectancy correlation:**

This figure combines the information from figure V.14b and V.15d. Next to the log 2 fold change heat map (vs control), it plots the correlation coefficient for each metabolite with life expectancy and the significance level of the associated p-value. It can be seen that metabolite candidates, which are associated with the inter-strain metabolite pattern also significantly correlate with life expectancy. Many are lipids.

tions and life expectancy. Furthermore, most of the metabolite candidates significantly correlated with life expectancy are lipids.

V.3.4.6. Particular phospholipid classes correlate with life expectancy

As shown above, many different metabolite classes are directly correlated to longevity. As the molecular compositions for all of the metabolite candidates are known, database searches can provide putative assignments for each correlation readily. However, as each of those correlations could tell its own story and generate hypotheses, looking at individual metabolite candidates would exceed the scope of this thesis. Thus, it was decided to continue with "big picture" analyses, focusing on metabolite classes, rather than individual

compounds.

Lipids were clearly the most significant metabolite class. Even in the most conservative scenario (figure V.15d), many late retention time features have significant correlation coefficients. Within those, two main clusters exist. At around 800 seconds, there are several features, which are negatively correlated. At around 900 seconds, however, many features have a positive coefficient of correlation of up to $R \approx \rho = 0.45$. When comparing the two correlation clusters with the atom-count plots from figure V.5, various sub clusters can be seen with different nitrogen, oxygen and phosphorus counts in particular. This is not surprising, as those elements differentiate between different lipid classes, for example between phospholipids and triglycerides. To find out more about the nature of those correlated clusters, an attempt was made to visualise the different lipid classes.

Figure V.17a shows a 2-D chromatogram of the lipid region, with different lipid groups highlighted. A lipid group was defined as metabolites with equal oxygen, nitrogen and phosphorus atom counts (table V.17b). These groups do not unambiguously correspond to biological lipid classes, such as those defined by LipidMaps [550], but are a good estimation. Some of the lipid classes were putatively assigned. The early retention time cluster consists of lipids carrying one fatty acid (FA), and the late retention time cluster are lipids with two fatty acids. In the second cluster, many lipids are subject to in-source fragmentation: the loss of the phosphocholine head group or one fatty acid can be observed. Thus, these signals are in the same mass range as the mono-fatty acid cluster, but may be discarded from further analysis. Figure V.17a also shows how the two correlation clusters (and some surrounding lipids) themselves readily divide into different subgroups. Those subgroups are lipid clusters of the same lipid class, but with different carbon and hydrogen counts, *i.e.* different FA chain lengths and degrees of saturation. Figure V.17c and V.17d show the positively and negatively correlated lipids, respectively.

All of the correlated lipid classes are glycerophosphocholines, also referred to as phosphatidylcholines (PC), as the loss of specific in-source fragments shows (see appendix figure C4.4). Those decreased in long-lived mutants have one FA conjugated to their glycerol backbone (monoacylglycerophosphocholines, monoAcPC), and the ones increased carry two FAs (diacylglycerophosphocholines, diAcPC). monoAcPC and diAcPC are both phosphatidylcholines, which have their FAs linked to the glycerol backbone via ester bonds.

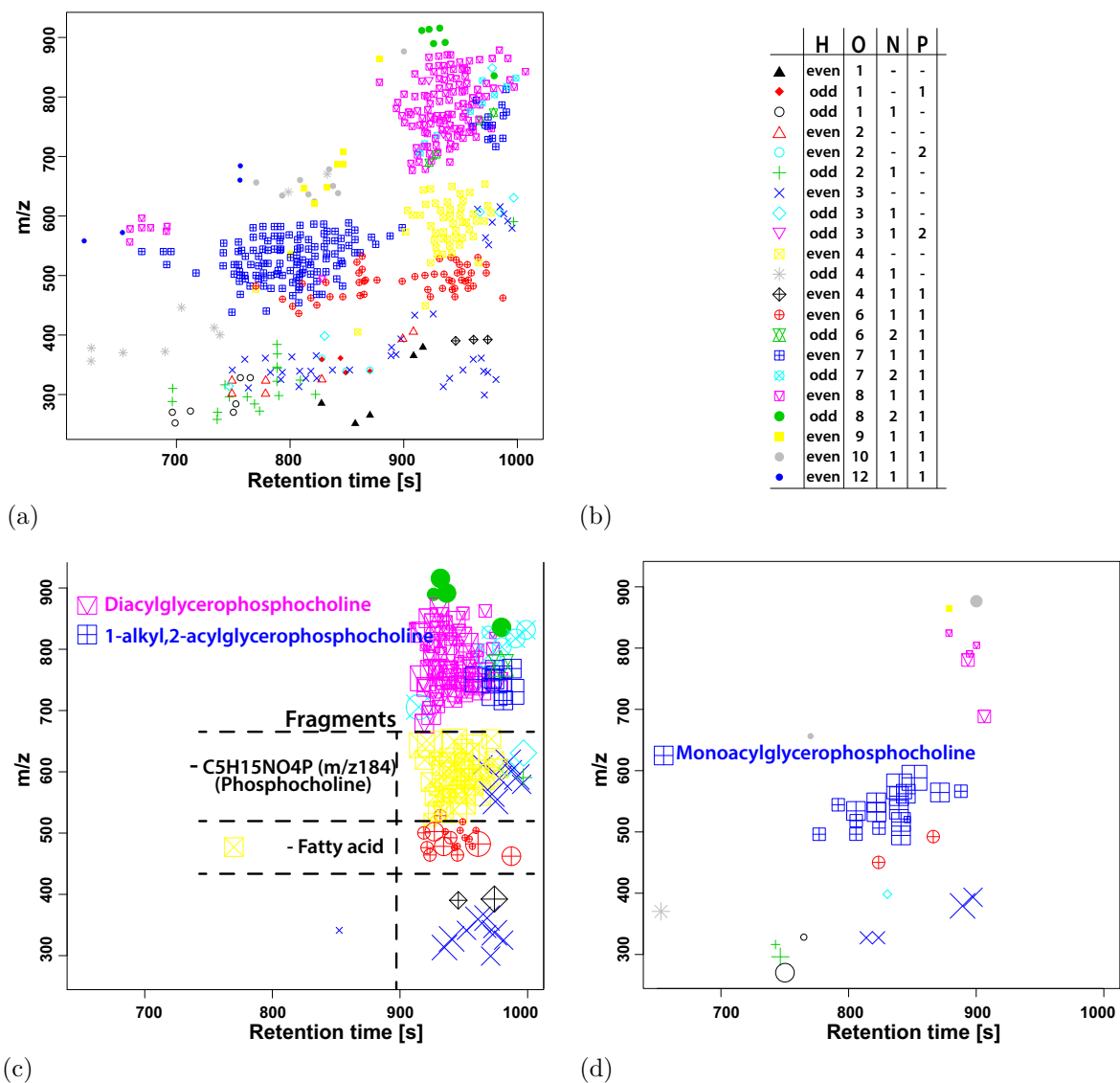


Figure V.17.: **Different lipid classes determine life expectancy:**

Figure V.17a groups all lipids in groups of common oxygen, nitrogen and phosphorous count (legend in figure V.17b). Major upregulated and downregulated lipid classes, together with their putative assignments, are shown in figure V.17c and figure V.17d, respectively.

Another upregulated group were 1-alkyl,2-acylglycerophosphocholines, which have one FA linked as an ester, the other one as an ether. In the second cluster, there are also a few dialkylglycerophosphocholines and sphingomyelins, which are lipids with their FAs linked via two ether bonds or via an amide bond, respectively. However, as the mono- and diacylglycerophosphocholines represent the lion's share of all correlated lipids ($n_{mono} = 119$, $n_{di} = 125$), only those were considered further.

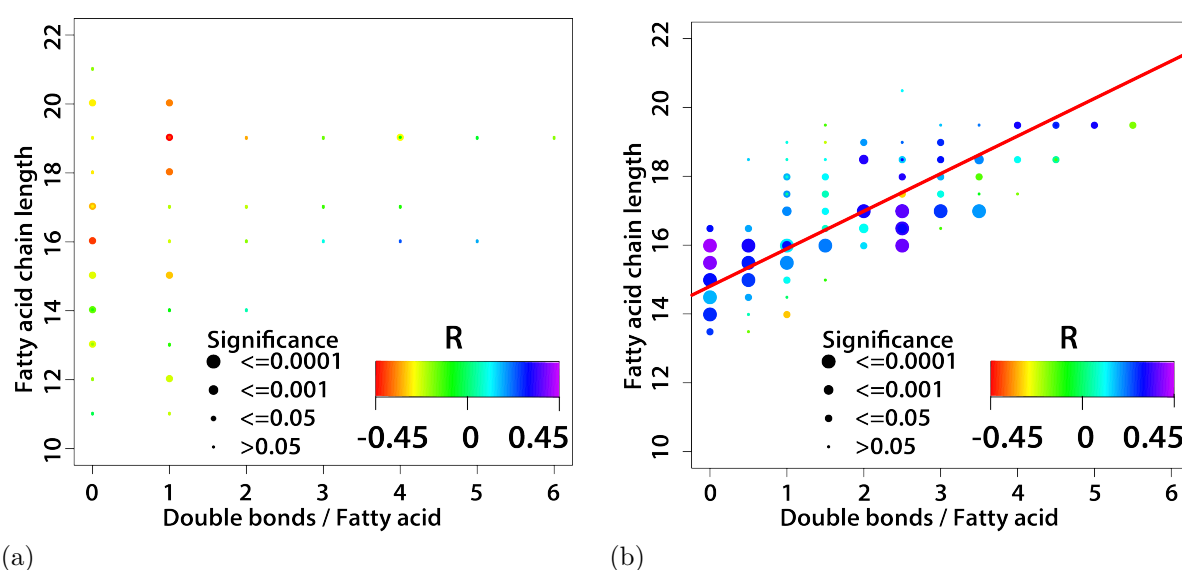


Figure V.18.: **Chain length, degree of saturation and correlation with life expectancy of mono- and diacylglycerophosphocholines:**

The degree of saturation, *i.e.* the average number of double bonds per fatty acid (FA) and the average carbon chain length is plotted for all monoacylglycerophosphocholines (monoAcPC) (figure V.18a) and diacylglycerophosphocholines (diAcPC) (figure V.18b). For better comparison, for the diAcPC, the number of double bonds and chain length is scaled to a per fatty acid, rather than per molecule basis. For the diAcPC a red median regression line (Tukey's median median) [581] is drawn between the significantly ($p \leq 0.05$) correlated metabolite candidates.

The oppositely correlated lipid classes were primarily differentiated with regards to the number of fatty acid tails per glycerophosphocholine backbone. However, they may also differ by the nature of the coupled FAs. Two main properties that characterise a FA are their carbon chain length and their degree of saturation. (The position of the double bonds may also be of interest, but was not considered here as it would require a different exper-

imental approach, such as MS-MS fragmentation studies or lipid fractionation, hydrolysis and subsequent GC-MS analysis.) Those two properties and the correlation coefficient and p-value from last chapter's correlation analysis (figure V.15d) are displayed in figure V.18.

Figure V.18a shows that all significantly ($p \leq 0.05$) decreased monoacylglycerophosphocholines carry saturated or monounsaturated fatty acids (MUFA). As seen in figure V.18b, the significantly ($p \leq 0.05$) increased diacylglycerophosphocholines are linked to fatty acids of various degrees of saturation. Their fatty acids can be polyunsaturated fatty acids (PUFA) with up to 5 double bonds. On average, the monoAcPC had higher fatty acid chain lengths than diAcPC lipids. For the significantly changed diAcPC, a direct correlation between chain length and number of double bonds is visible. Indeed, both values correlate ($\rho = 0.80$) significantly ($p < 0.0001$), according to a Spearman rank correlation test.

V.4. Discussion

V.4.1. A fair compromise between targeted and untargeted

Metabolomics [171] and metabonomics [444] were introduced as completely untargeted approaches, aiming to assess concentrations of all metabolites within a given matrix. Metabolomics was soon recognized as a powerful tool for hypothesis generation, and for functional genomic annotation [485, 440, 211]. However, a completely untargeted strategy also has drawbacks. Modern instruments produce so many data, it becomes impossible to annotate every molecule or even assess whether a peak is an actual metabolite. As shown, the vast majority of peaks are real signals, but are not biologically relevant. The risk of misleading findings increases with the number of detected signals.

For those reasons, targeted approaches, which try to assess a limited, but known subset of metabolites have become more popular. For LC-MS, two groups have developed this in particular: the Sauer [68] and Rabinowitz [29, 375] group from ETH and Princeton, respectively. Targeted methods are particularly powerful for hypothesis testing, as the activity of whole biological pathways [457, 168] can be reconstructed. This allows for more direct biochemical modelling approaches, rather than purely statistical ones. Targeted analysis requires some *a priori* decisions with regards to the metabolites. This, however, may miss biological information.

Here, an approach is demonstrated to annotate all biological peaks within a matrix and target them in the dataset. This method removes non-biological signals, and also removes redundancy within the biological signals; $M + n$ isotope peaks for example. Adducts could also potentially be removed, although this was not done here. Fragments, however, were annotated and removed from the data analysis. Removing redundancy not only reduces the risk of false positive assignments, it also improves the quality of multivariate models, by giving equal weight to each metabolite - rather than each signal. Thus, more liberal and potentially revealing modelling approaches may be used. After modelling relevant metabolites may even be identified more rapidly, as the molecular composition is already known. Unfortunately, one molecular composition may represent several metabolites. Thus, pathway based modelling approaches may not be feasible, without further identification. However, as demonstrated, metabolite subclasses, such as lipids, including their properties, such as number of carbon chain length or degree of saturation, can readily

be assessed and compared.

Despite all those benefits, the strategy also has certain drawbacks. Firstly, it requires the sample to be stable isotope enriched, which may exclude some common samples, such as biofluids. However, it may be possible to use an *E. coli* or yeast extract to assign common metabolites, within those matrices. Another potential problem is that only metabolites which were assigned in the first instance can be found in the dataset. Thus, if labelling is performed using the control samples, but a certain biomarker is only visible within one treatment condition, this molecule may be missed. However, it is questionable whether this risk outweighs the risk of a false assignment in an untargeted study or justifies the resources to produce a stable isotope version for each experimental condition, such as required by IROA (isotope ratio outlier analysis) [131, 69]. However, if an additional untargeted analysis of the data is performed, the true biological peaks may help with feature alignment or even improving the peak detection parameters in the first place. It may also be possible to predict the retention time [419, 117] and m/z of missing metabolites of a certain group. For example, the position of undetected members of a lipid class may be inferred by interpolating between the chain lengths and degree of saturation of annotated members of this class.

In the end, the choice of a method will depend on the application. The method described here is a good compromise between targeted and untargeted approaches. Ultimately, there can never be a truly untargeted method, as every analytical platform is selective towards certain compound classes [98].

V.4.2. An increased phospholipid pool links *C. elegans* longevity pathways

In this project, a large array of *C. elegans* mutants, which are long-lived due to modulation of various (independent) mechanisms of lifespan extension, were queried for their common metabolic changes. There was a clear set of commonly up and down regulated metabolite candidates with respect to controls, and most of those up/down regulated metabolite candidates had a significant correlation with life expectancy. The bulk of those potential longevity biomarkers are phospholipids. This is exciting, as recently research has found that many different core longevity mechanisms are tightly linked and dependent on lipid

metabolism itself [1, 346, 226]. In this study, two main lipid clusters could be identified. Glycerophosphocholine lipids linked to one fatty acid showed a negative correlation with longevity, and lipids with two fatty acids, showed a positive correlation. The downregulated lipids were monoglycerophosphocholines (lysophosphatidylcholines), but several different classes of 2 fatty acid lipids were upregulated. However, the majority of those were diacylglycerophosphocholines (phosphatidylcholines).

How are these lipid classes connected? Diacylglycerophosphocholines are a major constituent of the phospholipid bilayer and a hub of lipid synthesis pathways [158]. However, monoglycerophosphocholines (monoAcPC) and diacylglycerophosphocholines (diAcPC) can be converted into each other by fatty acid cleavage or transfer through phospholipase A and acylglycerophosphate acyltransferase, respectively. Thus, the simplest explanation for the observed strong increase of diAcPC and generally much less severe decrease of monoAcPC would be an increase of the diAcPC pool, depleting the monoAcPC reservoirs. This would also explain why in several independent studies [126, 183, 15] some long-lived mutants had a significant decrease of free glycerophosphocholine base - the precursor for monoAcPC. Indeed, Perez *et al.*[474] found that *daf-2* mutants, which had the most severe diAcPC/monoAcPC in-/decrease in this study, have an increased overall phospholipid pool (compared to wild type), with a significant contribution from *de-novo* synthesis. Thus, it seems that all mechanisms of lifespan extension share an elevated pool of phospholipids, and phosphatidylcholines in particular. But what is the mechanism behind this link?

Lipids are an organism's major long-term energy store. However, in *C. elegans*, there is no general correlation with total lipid storage and longevity. Germline as well as insulin-like signalling mutants have elevated fat stores [608, 474]. Dietary restricted mutants, which show the same general pattern in this study, are lean. Lipids, including certain phospholipids, act as specific signalling (second messenger) molecules [642]. However, in this study a non-specific increase of whole lipid classes, rather than individual (signalling) molecules, was observed. As phosphatidylcholines are the main constituent of biological membranes, maybe an increase in phospholipids coupled to specific fatty acids, *i.e.* the change of certain biophysical membrane properties, may prolong lifespan. Recently, Reis *et al.*[489] suggested that some of the longest lived ILS mutants, such as certain *daf-2* alleles, show an increase of more saturated fatty acids, and therefore are less prone to peroxidation. However, there is an agreement that the radical oxygen (ROS) theory is too

simplistic and fails to explain many mechanisms of lifespan extension [139], particularly as several modes of lifespan extension seem to depend on ROS or are even mediated by ROS [26, 268, 648]. Also, the superoxide dismutase mutant, which was analysed showed the general longevity pattern [589]. In this study, the degree of saturation of the significantly increased diacylglycerophosphocholines was analysed. There was no selectivity; saturated, mono saturated and even many polyunsaturated fatty acids were found to be conjugated. There was a correlation between the number of double bonds and fatty acid chain length, but this, however, is a known biochemical mechanism [563, 433] to maintain the physicochemical properties of the lipid membrane: decreased membrane fluidity due to increased fatty acid chain length can be offset by introducing more double bonds. Thus, overall, there seem to be no longevity effects due to changes in the physicochemical properties of the phospholipid bilayer itself. However, the phospholipid bilayer does not only make up the cell membrane. In fact, the cell membrane only constitutes two percent of a cell's phospholipid bilayer [7]. Phospholipids provide all means for compartmentalisation to the cell. Thus, more phospholipids will provide more compartmentalisation.

What does increased compartmentalisation have to do with longevity? Potentially, a lot. A third and nearly half of a cell's total phospholipid bilayer is used to form the cell's mitochondria and the Golgi apparatus plus endoplasmic reticulum, respectively. There is ever increasing evidence that these two organelles act downstream of many classical longevity (signalling) pathways to provide two key longevity effector mechanisms [371]: improved bioenergetic competency and mitochondrial biogenesis [66, 493, 26, 28], and an increased recycling of unused or damaged cell organelles and biomolecules by means of vesicular autophagy [383, 505]. Both processes require an increased amount of compartmentalisation and therefore phospholipids, as summarised by figure V.19. As recently shown, lipolysis also works by autophagy. Thus, the joint process of lipolysis and autophagy has been termed lipophagy [526]. Indeed, in most long-lived mutants, next to autophagy, a large scale remodelling of lipid pools, mainly a lipolysis of triglyceride stores, can be observed [458]. Dietary restriction [283, 576, 425], sirtuins [425, 426], insulin-like-signalling [576, 232, 608] (and therefore sensory mutations and DAF-16 degradation machinery), TOR [576], germline signalling [608, 345], decreased mitochondrial respiration [576, 649] and possible SOD downregulation all implicate or even depend on lipophagy for lifespan extension. The translation mutants, particularly *rsks-1* showed a metabolic phenotype

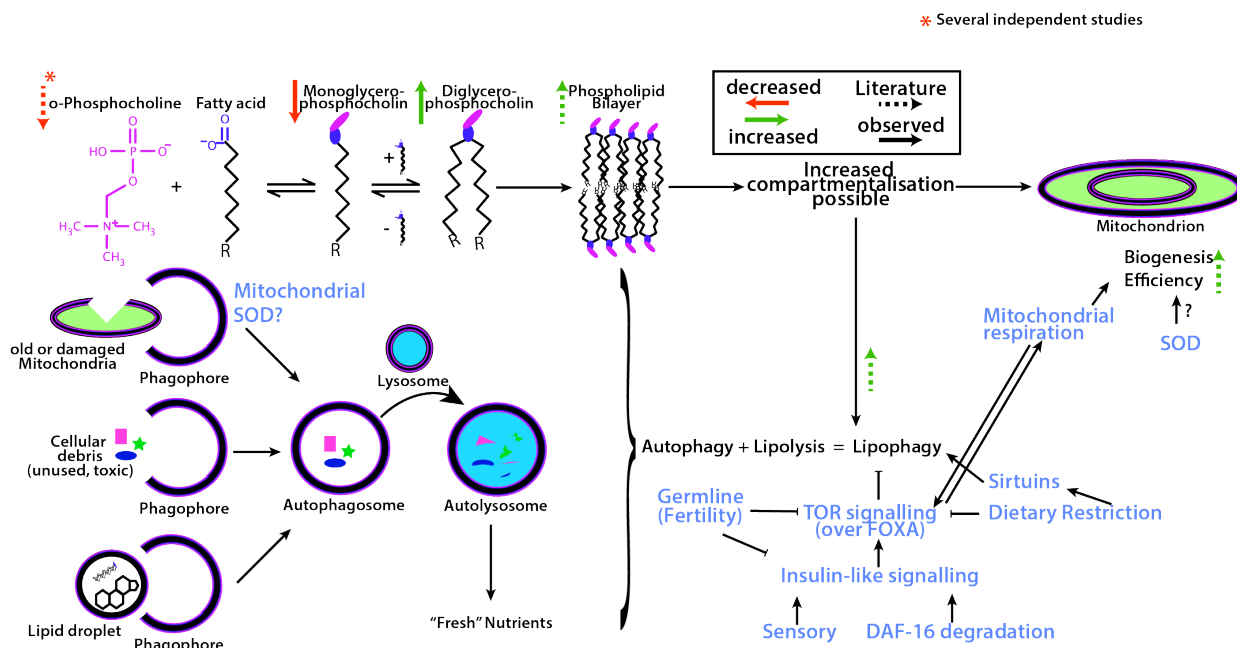


Figure V.19.: **Proposed model of longevity through increased compartmentalisation:**

Longevity pathways, which in literature have been shown to act through the corresponding effector mechanism are coloured in blue. This diagram is an over-simplification of the actual pathways themselves. Please also refer to the introductory figure V.1 and the referenced literature.

which was distinct, but slightly different from the common pattern. This confirms that they can perform lipophagy, but do not depend on it and there are additional effector mechanisms this model cannot account for. Autolipophagy is also a major survival mechanism, which is required for the extreme longevity of the dauer larval stage [410]. Thus, increased compartmentalisation may simply provide means to an organism, to utilise its resources more efficiently and renew its organelles, which effectively eliminates damage and leads to a "younger" phenotype. Decreased ageing and longevity would be the consequence.

V.4.3. The next steps

A simple model was proposed, which explains this experiment's results and is in agreement with all major lifespan extension pathways. However, a few mutants, such as the

hif-1 mutant, failed to show the common pattern. Perhaps this mutant simply did not extend longevity under the experimental conditions used. Indeed, for *hif-1*, which is temperature sensitive, 20°C leads to a modest increase in lifespan, if any [356]. Thus, while it was not feasible to perform longevity assays for all of the mutant strains in this thesis, ideally they would be repeated rather than relying on literature data. This would almost certainly improve the statistical relationships found, as exact lifespan extension values are notorious for varying slightly between labs.

In this thesis, only positively ionised metabolites which retained on reversed phase were analysed. For more comprehensive coverage, however, both ESI modes and different modes of chromatography should be used [588, 67]. This could easily be done in the future. It would be interesting to look at central carbon metabolism, and energy, redox and lipid biosynthesis precursors particularly. Energy and redox equivalents are the triggers for the AMP-kinase (AMPK) [20] and Sirtuin [270] lifespan extension mechanism, respectively. *daf-2* was shown to have an altered carbohydrate and amino acid metabolism [183] and mitochondrial mutants modulate small organic acid metabolism [75]. More detailed lipid analysis could also be performed. Unfortunately, due to the chosen extraction method [187, 186], triglycerides were not extracted. However, methods for *C. elegans* lipidomic analysis have been published [396], and again could easily be adopted in the future.

The workflow needs further development in general. For now, only molecular compositions are generated. Using MS/MS fragmentation, those compositions could be converted to actual metabolite (names). It would be easy to implement, to have automatically generated MS/MS target lists for all peaks identified as being of biological interest. The resulting fragment spectra could be matched against fragmentation databases [530, 251]. If a metabolite could not be matched within such a database, *in silico* fragmentation tools [632, 193] may be used. A new tool, LipidBlast, recently developed by Kind *et al.* [322] may be helpful to assign the fatty acids conjugated to lipids. In any case, putting actual names to metabolite candidates would enable a whole new dimension of analyses, such as pathway based modelling approaches.

Finally, the proposed model will require some biological validation. It would be interesting to see how lipids change over time. The challenges of such time-course ageing experiments will be discussed in the next chapter. Molecular methods may be used for hypothesis

testing. Particularly imaging approaches, like mass spectrometry based imaging [545], also discussed in the final chapter, may help to determine whether the compartmentalisation hypothesis is correct.

V.5. Conclusion

This chapter applied the method developed in the previous chapter to annotate *C. elegans* metabolites in a panel of long-lived mutant strains to find a common metabolic signature of longevity.

760 biological peaks were detected and assigned with molecular compositions. This equated to a data reduction of over 99.5 %. However, as shown later, this still includes (lipid) fragments and adducts. Those metabolite candidates were detected in a dataset of 24 long lived *C. elegans* mutants with various mechanistic bases for lifespan extension. It was shown that metabolite candidates can be used to align all samples of the dataset; at a step where standard algorithms failed. Also, this chapter provided a good example of how processing parameters can be chosen on a rational, experimental basis, rather than by best guess or relying on default values only. Univariate and multivariate models were built to find a common signature of longevity: lipids were a key metabolite group, with a sharp increase in phosphatidylcholine concentrations across all strains. This led to the formulation of a model, in which increased phospholipids provide the cell with additional means of compartmentalisation. This compartmentalisation is used to invoke a common effector mechanism of many of the longevity pathways - lipophagy. Also, it allows for increased mitochondrial biogenesis. It may be concluded, that by providing the structural means to perform recycling of old, unused and damaged cellular material, an increase of the phosphatidylcholine pool enables an organism to rejuvenate itself and become long-lived.

VI.. General discussion, future technologies and perspectives

The overall aim of this PhD was to develop methods for metabolomic analysis of the model organism *Caenorhabditis elegans*, that could then be applied to questions of ageing. A specific focus was to combine state-of-the-art analytical chemistry and computational data-processing approaches. Ultimately, several software tools and workflows were developed which could be useful not only for *C. elegans* ageing research, but the wider metabolomics community, particularly for work with model organisms. The final two experimental chapters showed how powerful the combination of analytical and computational approaches can be, when applied to important biological questions.

In this chapter, an overview of the major findings will be given, and two new approaches presented, which have a high future potential in this area. The relevance of this thesis to the current state of metabolomics and ageing research will be discussed.

VI.1. Summary of the key results

The first results chapter laid the basis of this PhD by developing an extraction protocol which maximises coverage, yield and reproducibility, and is compatible with multiple analytical platforms. As well as deciding on a final optimized method, 80 % methanol extraction using a bead-beater, this chapter also provided a first snapshot of the *C. elegans* metabolome on these three platforms. The chapter focussed on the (often neglected) tissue disruption requirement for metabolite extraction, and also found that the solvent system used had a large impact on the final result. A generally applicable key finding was that certain extraction protocols may appear to work well on one, but not on another analytical platform. Thus, it was recommended to, ideally, always benchmark protocols

across multiple analytical platforms. Finally, it also identified limits and shortcomings of the three analytical platforms for *C. elegans* metabolomics, which were addressed in the next two chapters by extensive method development.

In the second results chapter, simple improvements to methods were tested, which aimed to address limitations of the the used workflow or NMR and GC-MS metabolite profiling, with a particular focus on improving metabolite annotation and assignment.

A method to automatically annotate HSQC spectra was developed. For sensitivity improvement, the *C. elegans*, worms were fully ^{13}C enriched. Adoption of a new heteronuclear 2D pulse sequence, widely used in structural biology for labeled proteins, yet untested for small molecule analysis of stable isotope enriched samples (ct-HSQC), reduced carbon-carbon splitting. Removal of this "artefact" meant that it was feasible to use the labeled samples to improve sensitivity, yet still peak pick the spectra and to match against the online databases. This gave quick and comprehensive automatic assignment of many of the small molecules of this extract. These were "pre-validated" in the sense that many of them had already been assigned manually in the first chapter, but it demonstrates proof-of-principle that the approach worked, while giving greater coverage of the *C. elegans* metabolome.

Next, a LC-NMR ion-pairing strategy was developed. It was found that deuteration of the ion pairing reagent does not negatively impact separation performance, but improved spectral quality by silencing of its otherwise NMR visible protons. In general, ion-pair chromatography separated several compounds from the NMR visible *C. elegans* metabolome and revealed metabolites obscured in conventional 1-D tube-based NMR. It was also shown that ion-pairing LC-MS is useful to annotate compounds which are difficult to assign in a 1-D or even 2-D spectrum, such as ATP, ADP and AMP.

The final method described in the second results chapter, was a approach to complement the current library-based GC-MS workflow with an untargeted approach. It was shown that alignment and overlay of the total ion chromatograms (TIC) of different sample groups is an effective way to determine differences in metabolite concentration without any *a priori* knowledge about the identity of the metabolites. Two additional metabolites were detected that weren't identified by the standard workflow, and were significantly different between

short-lived and long-lived *C. elegans* mutants. Interestingly, these two metabolites, 2-aminoadipic acid and 2-hydroxyglutaric acid, have already received significant attention in the (ageing) literature. The new data could be used to generate hypotheses for future validation.

In the third results chapter, an isotopic labelling method was developed, which addresses two current major problems of LC-MS metabolomics: data reduction and compound annotation. Complete *ab initio* molecular formula generation based on accurate mass only is unrealistic with current technology. However, it was demonstrated that determining the carbon, nitrogen and sulphur atom-count by isotopic labelling means that all molecular composition up to 1000 Da can be resolved unambiguously. This led to the formulation of a triple stable isotope labelling protocol in *E. coli*, and also *C. elegans*; the first example of triple labelling in an animal, although plants have previously been triply labelled. A software package was developed to analyse the data resulting from LC-MS runs of triply labelled samples, to detect all peaks of biological origin, and annotate them with their molecular composition (formula). Using *E. coli* extracts, it was shown that this software algorithm works for different chromatographic conditions and for both positive and negative electrospray ionisation. Most surprisingly, more than 99.9 % of all LC-MS detected peaks turned out to be non-biological, and hence not metabolites.

The final results chapter used this isotopic labelling method to analyse a set of 24 long-lived *C. elegans* mutant strains. First, all biological peaks were annotated with molecular compositions. In a second step, those target peaks were detected, aligned and integrated in the mutant dataset. It was shown, that those target peaks can be used to rationally optimise the algorithm's parameters and also help to retention time align the samples; a procedure which did not succeed using other contemporary software packages. Finally, there was a metabolic signature of long life, *i.e.* a set of metabolites which were associated with longevity across the set of mutants analysed. The main metabolites correlated with longevity were lipids. The calculated molecular compositions could not only be used to readily annotate different lipid classes, but also to calculate their properties, such as length and degree of saturation of their fatty acid chains. The long-life signature turned out to be because of a general increase of the glycerophosphocholines lipid pool. This was used to propose the "compartmentalisation hypothesis". More glycerophosphocholine lipids are present because they are needed to build sub-cellular membranes for compartmentalisa-

tion, which form the basis of several recently described longevity effector mechanisms: mitochondrial biogenesis, lipid hydrolysis and autophagy. Thus, it seems that most longevity pathways share common effector mechanisms, which are linked by a increased requirement for compartmentalisation and therefore glycerophospholipids.

VI.2. Two key problems and and two promising key technologies

This thesis provides solutions to several problems encountered in metabolomics. However, there still remain many questions that require further investigation or improvements in methodology. Throughout this thesis, suggestions for improvements of the proposed methods or comments on required biological follow up experiments were made. Rather than providing another list of hypothetical issues that should be tackled, two promising new tools that emerged from this work, and their potential to help answer two key questions, will be discussed. Development of both methods started during this PhD and the proof of concepts were published [84, 185]. Even though they still require a substantial development and testing to become routine tools, they both have high potential to help elucidate the metabolic factors that affect ageing and longevity in *C. elegans*.

VI.2.1. The problem of measuring ageing: "The worm sorter"

This thesis's title is "Analytical and computational methods towards a metabolic model of ageing in *Caenorhabditis elegans*". The first 3 results chapters were concerned with developing those methods. The last chapter then used those methods to propose a model for the metabolic basis of longevity in *C. elegans*. Whereas longevity is tightly connected to ageing, metabolic changes throughout the ageing process itself must also be investigated. However, during this PhD it became apparent that several problems would have to be solved in order to obtain the highest quality data. A key problem is the mixing of generations. After 3.5 days of development, an adult hermaphrodite worm lays up to 300 eggs, and after another 3.5 days these will produce offspring themselves. By then, the number of offspring will have superseded the starting generation by more than two orders of magnitude. This means that it is necessary to separate adults from offspring, otherwise

synchronized populations rapidly become a mix of adults and juvenile worms.

What alternatives are there to a "worm sorter". Worms can be manipulated on a individual basis and separated from their offspring by transferring them to new agar plates. However, as this requires daily "picking" of worms, it is not a viable option for the thousands of worms required for a single metabolomics sample, with the current analytical workflow. Another method maintains a large population of age synchronised worms by using mutants which are sterile when maintained at higher temperatures [287, 377, 432, 227], similar to the fertility mutants used in the longevity mutant screen. However, as such a condition is far from normal physiology, impacts life expectancy itself, and as shown in the last chapter produces a distinct metabolic phenotype, this is not an option either. Another common method is the addition of FUdR (5-fluoro-2-deoxyuridine) to the medium [224, 353, 340]. FUdR is a mitosis inhibitor. As *C. elegans* cells stop dividing after the larval stage, FUdR will arrest all progeny at larval stage but keep their parents alive. Until recently, FUdR was thought to be rather inert with regards to adult worm physiology. However, it can significantly extend the lifespan of certain mutant strains [5, 590]. For example, life expectancy of the *gas-1* mutant, analysed in the GC-MS alignment chapter, can nearly be doubled by the use of FUdR. This effect is even dose dependant [590]. Davies *et al.*[125] also showed that FUdR has a significant impact on the metabolic phenotype. It was even able to suppress metabolic longevity responses of *daf-2* mutants; the mutation which showed the strongest metabolic phenotype in the mutant screen. For those reasons FUdR cannot serve as a solution to the "offspring problem".

Therefore, a microfluidics based solution - "The worm sorter" - was developed in collaboration with Xavier Casadevall i Solvas from the deMello lab [84]. This is a small microfluidics based device to separate adult worms from their offspring. Microfluidics is an emerging technology, aiming to produce low-cost [399] devices which can manipulate fluids on a nanoscale, with applications in chemical synthesis, biological analysis, diagnostics and many more fields [623]. It has already been used to study live animals like *C. elegans*[343, 104, 381, 266]. Even some worm sorting systems had been developed. However, they all rely on high tech instrumentation, involving image acquisition of individual worms, computational classification and the actuation of micro-valves to sort them [503, 100, 267, 116, 507]. In this project, a small, low-cost device was developed (over several generations), which allows to separate adult *C. elegans* from their offspring - by a

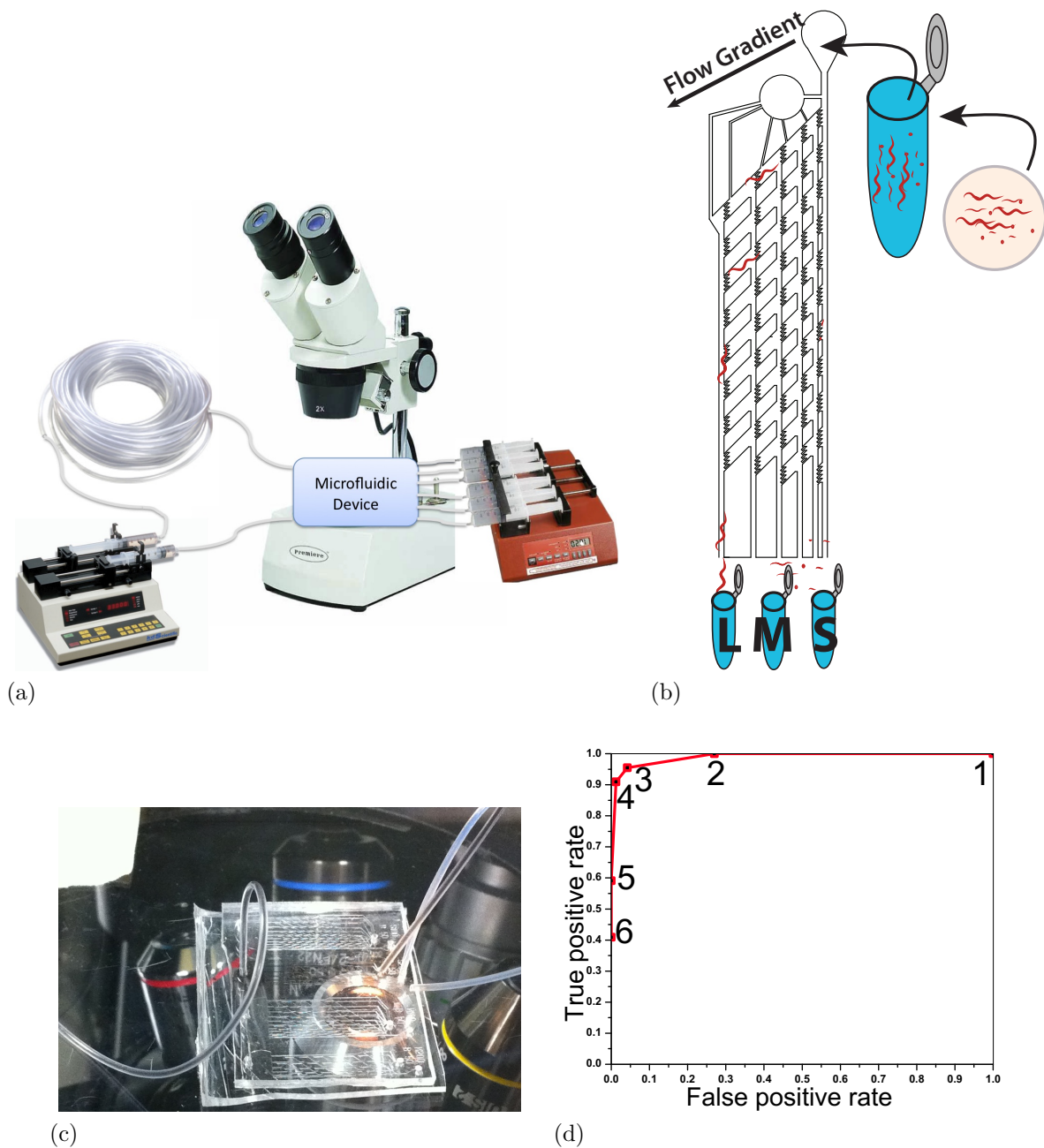


Figure VI.1.: **The "Worm sorter" may solve the "offspring problem":**

The microfluidic device (figure VI.1c) can sort adult worms from eggs, small and medium worms. It only requires a syringe pump to apply a flow gradient and another syringe pump (or Eppendorf tubes) to collect the sorted fractions (figure VI.1a). The sorting is performed by a smart maze, which redirects adult worms to take a different path than smaller individuals, which are simply flushed out through another exit by the applied flow gradient (figure VI.1b). The ROC curve in figure VI.1d shows that a high selectivity can be achieved, depending from which output the worms are sampled.

passive mechanism. Figure VI.1 shows the setup of this device. It only requires a syringe pump, but a microscope is useful during the setup phase (figure VI.1a). The chip (figure VI.1c) contains a smart maze of small channels with six outlets. The syringe pump creates a flow velocity gradient across the chip. When a mixed population is introduced into the chip, large worms are redirected by the maze to swim to the opposite side of the chip, whereas small worms and eggs are directly washed out by the flow (figure VI.1c). Thus, adult worms can be collected from outlet five and six, offspring (larvae and eggs) from outlet one and two. As the ROC (figure VI.1d) curve shows, the sorting is highly specific. In the latest generation devices, 94 % of all adults could be recovered with only 0.2% of larval contamination and a throughput of 200 worms per minute and chip. However, this approach is fully scalable by using chips in parallel or series to improve throughput or separation efficiency, respectively.

Unfortunately, this device is still in an experimental phase, and its operation and setup is still not straightforward. However, as the development is being continued by the deMello lab, one day, it will hopefully provide a superior alternative to all other solutions of the "offspring problem". This would allow time course experiments to analyse the impact of metabolism throughout the actual ageing process of a synchronous *C. elegans* population.

VI.2.2. Biology is spatial: TOF-SIMS

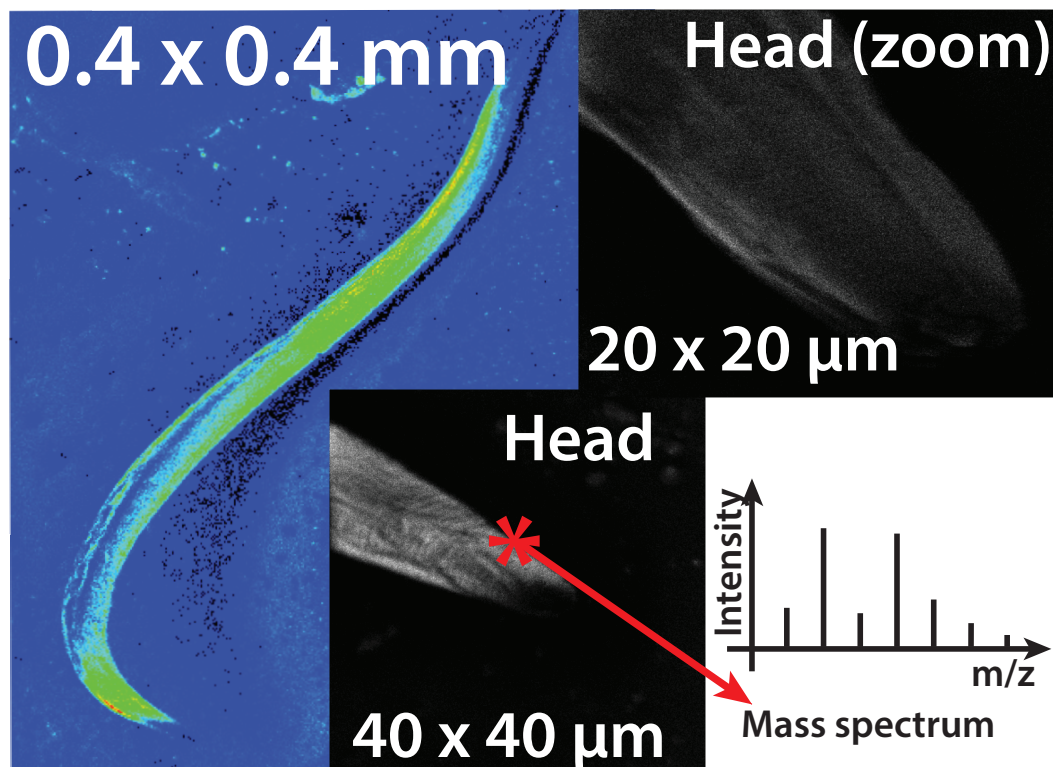
Paradoxically, a model of increased compartmentalisation was proposed on the basis of an extract, generated from thousands of worms, each with a thousand cells, which may have thousands and millions of compartments themselves. Averaging over thousands of individual systems, to provide an integrated phenotype is both a strength and the weakness of metabolomics [441]. However, validation of this hypothesis may require more specific techniques, which are able to show an increase in compartmentalisation, such as an increased autophagy and mitochondrial biogenesis.

As those processes are all at an upper nanometre scale [650], this stretches the limits of classic light microscopic techniques and confocal or even electron microscopy would be the preferred method [222]. Microscopy in general, however, does not provide chemical information. Thus, mass spectrometry based imaging is becoming an increasingly popular technique [545, 233, 447, 466]. Many different techniques exist, with matrix-assisted

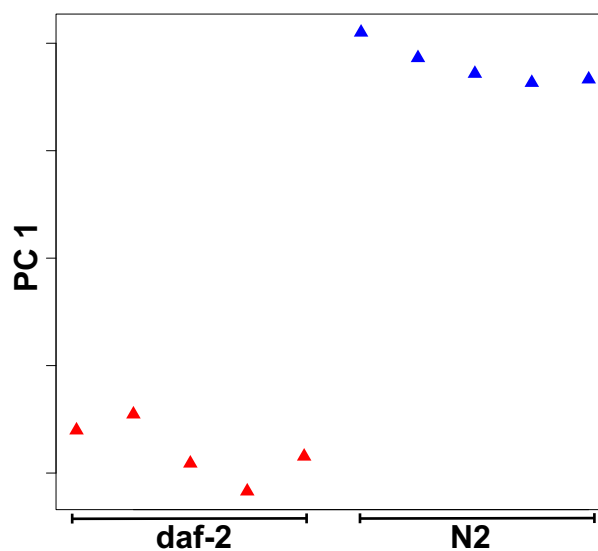
laser desorption ionisation (MALDI) probably being the most prominent. MALDI, however, works on a scale of several tens of micrometers [13, 112, 448] and therefore is not able to obtain cellular or sub cellular information from *C. elegans* [318]. Another technique, secondary ion mass spectrometry (SIMS) can be considered a cross between a mass spectrometer and an electron microscope [393], even with high mass accuracy capabilities (ToF-SIMS) [209]. Therefore SIMS should operate at the right length scales to visualise an increase in compartmentalisation. Indeed, ToF-SIMS was shown to be able to image a phospholipid bilayer with a lateral resolution of 100 nm or less [215]. In fact, lipids, with cholesterol and phospholipids in particular, are a molecular class which allows for very sensitive ToF-SIMS analyses [528, 406, 178]; even quantitatively [459]. It was even demonstrated that different lipid chain lengths can be discriminated [491]. Also, not possible with any of the other methods, ToF-SIMS can perform depth profiling. By abrogating the specimen's upper layer with the ion gun and subsequent imaging, ToF-SIMS can even produce 3D chemical images of single cells [177, 59]. Also, by using stable isotope labels the spatial dynamic of biomolecules inside single cells can be studied [543].

To evaluate the opportunity of using ToF-SIMS for spatial *C. elegans* metabolomics, a pilot project, together with Sarah Fearn and David McPhail from the department of materials science at Imperial College was initiated. Some of the already published results [185] are shown in figure VI.2. As figure VI.2a shows, ToF-SIMS imaging of worms, fixed by freeze-drying [389], was able to prize out morphological structures, which otherwise are only visible by electron microscopy. Unsupervised principal component analysis was also able to discriminate between the extracts of a wild type and *daf-2* mutant (figure VI.2b). In another experiment, it was possible to discriminate between different lipid classes (not shown).

However, in order to be able to use ToF-SIMS to answer biological questions, it is necessary to optimise sample preparation, in order to be certain that observations are not due to sample preparation artefacts [389]. Classic techniques for producing thin sections for EM involve resin embedding, which preserves macromolecular ultrastructure but completely destroys metabolic information. Cryosectioning approaches are challenging, even simply being to locate an individual worm at the correct orientation. However, ToF-SIMS promises to be a promising tool of great value to bring a spatial component to *C. elegans* metabolomics. Some of the key biochemical phenomena underlying ageing, such as



(a)



(b)

Figure VI.2.: **ToF-SIMS imaging mass spectrometry:**

ToF-SIMS can perform mass spectrometry based imaging, but at a electron microscopy like resolution. Figure VI.2a shows an imaged *C. elegans* larva, with sections of the head magnified. Morphological structures are clearly visible. However, these are false colour images of the total ion count. In fact, each pixel contains a complete high resolution mass spectrum. Figure VI.2b shows that a principal component analysis can discriminate between a ToF-SIMS acquired extract of a *daf-2* mutant and wild type. All data were acquired using a Bi^{3+} cluster liquid metal beam [577]

autophagy, cannot be fully understood without the spatial component [346, 458]. Indeed, not only for ageing but for many other kinds of biological phenomena that could be studied by metabolomics in *C. elegans*, imaging mass spectrometry, and other approaches such as Raman microscopy, can provide information not available any other way.

VI.3. Current issues

This section puts some of the key contributions of this PhD into the wider context of metabolomics and (*C. elegans*) ageing research.

VI.3.1. Metabolomics

In a comprehensive review, Dunn *et al.*[143] discuss "growing pains" in the field of metabolomics. The number one pain is "Chemical identification of metabolites"; methods are needed to automatically assign all metabolites contained inside a sample, rather than manually identifying metabolites of biological interest. Exactly this issue has been addressed by some methods proposed in this thesis. The ^{13}C ct-HSQC annotation workflow provides such a tool for NMR, albeit limited to metabolites present in the spectral database; which in this thesis was a fair amount of all resonances. Putative assignments will obviously require manual validation to cut down false positives. The other method, which works without any *a priori* metabolite information, is the isotope annotation workflow. This method has many applications, as shown in the last chapter. Unfortunately, this algorithm cannot deliver metabolite IDs. However, given the recent progress in *in silico* fragmentation tools [632, 193, 322], reliable automatic generation of metabolite IDs may not be far off. The analysis of the lipid classes in the longevity mutants was a demonstration of the type of analyses that could be performed, if all metabolites were annotated in each single study.

Another issue, albeit rarely mentioned, is the correct choice of parameters. Metabolomics is very dependent on computational tools, and correct parameter settings are crucial to its output and therefore success [142, 62]. Whereas this goes without saying for many wet lab procedures, such as actual sample acquisition, it often seems to be ignored at the sample processing stage. There are many possible reasons: not understanding the importance of parameter settings; uncertainty about how to optimise them; or even lack of access to these parameters, particularly with modern "push button" solutions on the rise. While

such solutions may outperform inexperienced users, they cannot outperform experienced operators and also bear a risk of producing suboptimal output. In this thesis, no parameter was left at default value. An attempt was made, to rationally find the optimum for each processing parameter. For example, it was found that certain peak detection parameters are even selective towards biological peaks. Even though those optimisations were facilitated by the fact that all code was developed or optimised by the user, such a rational search for the best variables should be encouraged within the community. People must realise, that correct parameter choices may contribute towards the quality of the result on a similar scale, as good practice sample quality control measures [511] and rigorous statistical data handling [61].

VI.3.2. (*C. elegans*) ageing research

Over the last two decades, much work has been done to elucidate pathways that affect lifespan. Major labs have worked on different parts of those different signalling network. Also, more work has been done on the signal transduction, rather than the downstream effector mechanisms, which ultimately lead to the lifespan extension. However, in the last few years, there has been a push to identify these mechanisms, such as the unfolded protein response [147], lipophagy [345] or mitochondrial biogenesis [26]. Unsurprisingly, most of the major longevity signalling cascades, often studied in isolation, ultimately feed into these few effector mechanisms. A common endpoint, downstream of various (independent) pathways, can be observed in the form of a common metabolic signature of the longevity mutants. This demonstrates that metabolomics, as the most downstream and integrated endpoing of an organism's gene, transcript, protein, metabolite cascade, was an appropriate tool to identify a common signature. Providing this integrated phenotype, is a key strength of metabolomics, and perhaps the more upstream or more specific "omics", such as transcriptomics or proteomics, may have not succeeded.

But, as metabolomics is unspecific, those other, more specific "omics" will be required to validate the proposed compartmentalisation model. Obviously, as worms are a simple model organism, this hypothesis must also be validated in other, more complex mammalian model systems, such as mice. Ultimately, it also has to be tested in humans, which for any scientific model is the most challenging task. However, very recently a clinical study on long-term dietary restriction in humans showed an up regulation of autophagic activity

and mitochondrial biogenesis and generally "younger" transcription profiles [411]. This is in absolute agreement with the model proposed here, which therefore may be applicable to human longevity and ageing; particularly as the molecular mechanisms this model relies on are very conserved.

Some of the mutations studied were in genes which, under physiological conditions, are modulated by the environment, such as the sensory mutants, the dietary restriction mutant and even the insulin like signalling mutants. This implies that genetic and environmental influences jointly affect the level of effector mechanisms. Thus, the same lifespan extension may be reached by environmental intervention or genetic predisposition. If this was also true for humans, the presence of common gene-environment longevity effector mechanisms could provide a molecular explanation, for the increase of life expectancy, observed in developed countries. It would also mean, that longevity may be modulated by conscious lifestyle decisions.

VI.4. Conclusion

In summary, combining methods from different fields, such as chemistry, biology and computing, can address questions that are difficult to answer within a single discipline.

The methods developed in this PhD thesis have addressed current needs for metabolomics in *C. elegans* and have also provided new insights into how different mechanisms of lifespan modulation may be linked on a molecular level.

Bibliography

- [1] D. Ackerman and D. Gems. The mystery of *C. elegans* aging: an emerging role for fat. Distant parallels between *C. elegans* aging and metabolic syndrome? *BioEssays*, 34(6):466–471, June 2012.
- [2] F. M. Afendi, T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, S. Ikeda, H. Takahashi, M. Altaf-Ul-Amin, L. K. Darusman, K. Saito, and S. Kanaya. KNAPSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant & Cell Physiology*, 53(2):e1, Feb. 2012.
- [3] M. Aickin and H. Gensler. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *American Journal of Public Health*, 86(5):726–728, May 1996.
- [4] M. Ailion, T. Inoue, C. I. Weaver, R. W. Holdcraft, and J. H. Thomas. Neurosecretory control of aging in *Caenorhabditis elegans*. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 96(13):7394–7397, June 1999.
- [5] L. Aitlhadj and S. R. Sturzenbaum. The use of FUdR can cause prolonged longevity in mutant nematodes. *Mechanisms Of Ageing And Development*, 131(5):364–365, May 2010.
- [6] K. Albert. *On-line LC-NMR and related techniques*. J. Wiley & Sons, New York, 2002.
- [7] B. Alberts. *Molecular Biology of the Cell: Reference Edition*. Molecular Biology of the Cell: Reference Edition. Garland Science, 2008.
- [8] J. Alcedo and C. Kenyon. Regulation of *C. elegans* longevity by specific gustatory and olfactory neurons. *Neuron*, 41(1):45–55, Jan. 2004.
- [9] S. G. Alikhanov. A new impulse technique for ion mass measurements. *Soviet Phys JETP*, 4:452–453, 1957.
- [10] J. W. Allwood and R. Goodacre. An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis*, 21(1):33–47, Jan. 2010.
- [11] E. Alm, R. J. O. Torgrip, K. M. Aberg, I. Schuppe-Koistinen, and J. Lindberg. Time-resolved biomarker discovery in H-1-NMR data using generalized fuzzy Hough

- transform alignment and parallel factor analysis. *Analytical and Bioanalytical Chemistry*, 396(5):1681–1689, 2010.
- [12] A. J. Alpert. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *International Journal of Pharmaceutics*, 499:177–196, Jan. 1990.
- [13] A. F. M. Altelaar, S. L. Luxembourg, L. A. McDonnell, S. R. Piersma, and R. M. A. Heeren. Imaging mass spectrometry at cellular length scales. *Nature Protocols*, 2(5):1185–1196, 2007.
- [14] T. Altman, M. Travers, A. Kothari, R. Caspi, and P. D. Karp. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, 14(1):112, 2013.
- [15] Y. J. An, W. J. Xu, X. Jin, H. Wen, H. Kim, J. Lee, and S. Park. Metabotyping of the *C. elegans* sir-2.1 Mutant Using in Vivo Labeling and (13)C-Heteronuclear Multidimensional NMR Metabolomics. *ACS Chemical Biology*, 7(12):2012–2018, Oct. 2012.
- [16] P. E. Anderson, D. A. Mahle, T. E. Doom, N. V. Reo, N. J. DelRaso, and M. L. Raymer. Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. *Metabolomics*, 7(2):179–190, 2011.
- [17] T. M. Annesley. Ion suppression in mass spectrometry. *Clinical Chemistry*, 49(7):1041–1044, July 2003.
- [18] J.-P. Antignac, K. de Wasch, F. Monteau, H. De Brabander, F. Andre, and B. Le Bizec. The ion suppression phenomenon in liquid chromatography–mass spectrometry and its consequences in the field of residue analysis. *Analytica Chimica Acta*, 529(1):129–136, 2005.
- [19] J. Apfeld and C. Kenyon. Regulation of lifespan by sensory perception in *Caenorhabditis elegans*. *Nature*, 402(6763):804–809, Dec. 1999.
- [20] J. Apfeld, G. O’Connor, T. McDonagh, P. S. DiStefano, and R. Curtis. The AMP-activated protein kinase AAK-2 links energy levels and insulin-like signals to lifespan in *C. elegans*. *Genes & Development*, 18(24):3004–3009, 2004.
- [21] N. Arantes-Oliveira, J. Apfeld, A. Dillin, and C. Kenyon. Regulation of lifespan by germ-line stem cells in *Caenorhabditis elegans*. *Science (New York, NY)*, 295(5554):502–505, 2002.
- [22] N. Arantes-Oliveira, J. R. Berman, and C. Kenyon. Healthy animals with extreme longevity. *Science (New York, NY)*, 302(5645):611, Oct. 2003.

- [23] H. J. Atherton, O. A. H. Jones, S. Malik, E. A. Miska, and J. L. Griffin. A comparative metabolomic study of NHR-49 in *Caenorhabditis elegans* and PPAR-alpha in the mouse. *FEBS Letters*, 582(12):1661–1666, May 2008.
- [24] P. Atkins and J. de Paula. *Atkins' Physical Chemistry*. OUP Oxford, 2009.
- [25] P. Ausloos, C. L. Clifton, S. G. Lias, A. I. Mikaya, S. E. Stein, D. V. Tchekhovskoi, O. D. Sparkman, V. Zaikin, and D. Zhu. The critical evaluation of a comprehensive mass spectral library. *Journal Of The American Society For Mass Spectrometry*, 10(4):287–299, Apr. 1999.
- [26] A. B Hwang, D.-E. Jeong, and S.-J. Lee. Mitochondria and Organismal Longevity. *Current Genomics*, 13(7):519–532, Sept. 2012.
- [27] V. I. Babushok, P. J. Linstrom, J. J. Reed, I. G. Zenkevich, R. L. Brown, W. G. Mallard, and S. E. Stein. Development of a database of gas chromatographic retention properties of organic compounds. *Journal Of Chromatography A*, 1157(1):414–421, 2007.
- [28] P. Back, B. P. Braeckman, and F. Matthijssens. ROS in Aging *Caenorhabditis elegans*: Damage or Signaling? *Oxidative Medicine and Cellular Longevity*, 2012:14, 2012.
- [29] S. U. Bajad, W. Lu, E. H. Kimball, J. Yuan, C. Peterson, and J. D. Rabinowitz. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *Journal Of Chromatography A*, 1125(1):76–88, Aug. 2006.
- [30] K. K. Banerjee, C. Ayyub, S. Z. Ali, V. Mandot, N. G. Prasad, and U. Kolthur-Seetharam. dSir2 in the Adult Fat Body, but Not in Muscles, Regulates Life Span in a Diet-Dependent Manner. *Cell Reports*, 2(6):1485–1491, Dec. 2012.
- [31] R. Baran, B. P. Bowen, N. J. Bouskill, E. L. Brodie, S. M. Yannone, and T. R. Northen. Metabolite Identification in *Synechococcus* sp. PCC 7002 Using Untargeted Stable Isotope Assisted Metabolite Profiling. *Analytical Chemistry*, Oct. 2010.
- [32] M. Barber, R. S. Bordoli, R. D. Sedgwick, and A. N. Tyler. Fast atom bombardment of solids (FAB): A new ion source for mass spectrometry. *Journal of the Chemical Society, Chemical Communications*, 0(7):325–327, 1981.
- [33] G. F. Barker. *A text-book of elementary chemistry, theoretical and inorganic*. C.C. Chatfield & Co., 1874.
- [34] T. M. Bass, D. Weinkove, K. Houthoofd, D. Gems, and L. Partridge. Effects of resveratrol on lifespan in *Drosophila melanogaster* and *Caenorhabditis elegans*. *Mechanisms Of Ageing And Development*, 128(10):546–552, Oct. 2007.

- [35] M. Bauer, A. Bertario, G. Boccardi, X. Fontaine, R. Rao, and D. Verrier. Reproducibility of ^1H -NMR integrals: a collaborative study. *Journal Of Pharmaceutical And Biomedical Analysis*, 17(3):419–425, July 1998.
- [36] O. Beckonert, M. Coen, H. C. Keun, Y. Wang, T. M. D. Ebbels, E. Holmes, J. C. Lindon, and J. K. Nicholson. High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. *Nature Protocols*, 5(6):1019–1032, May 2010.
- [37] O. Beckonert, H. C. Keun, T. M. D. Ebbels, J. Bundy, E. Holmes, J. C. Lindon, and J. K. Nicholson. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11):2692–2703, 2007.
- [38] V. Behrends, G. D. Tredwell, and J. G. Bundy. A software complement to AMDIS for processing GC-MS metabolomic data. *Analytical Biochemistry*, 415(2):206–208, Aug. 2011.
- [39] A. Beltran, M. Suarez, M. A. Rodríguez, and M. Vinaixa. Assessment of Compatibility between Extraction Methods for NMR- and LC/MS-Based Metabolomics - Analytical Chemistry (ACS Publications). *Analytical Chemistry*, 84(14):5838–5844, 2012.
- [40] R. Bender and S. Lange. Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology*, 54(4):343–349, Apr. 2001.
- [41] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1), 1995.
- [42] T. W. Bentley and R. Johnstone. Mechanism and structure in mass spectrometry: a comparison with other chemical processes. *Advances in Physical Organic Chemistry*, 1970.
- [43] H. Benton, E. Want, and T. Ebbels. Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data. *Bioinformatics (Oxford, England)*, 2010.
- [44] H. P. Benton, E. Want, H. C. Keun, A. Amberg, R. S. Plumb, F. Goldfain-Blanc, B. Walther, M. D. Reily, J. C. Lindon, E. Holmes, J. K. Nicholson, and T. M. D. Ebbels. Intra- and interlaboratory reproducibility of ultra performance liquid chromatography-time-of-flight mass spectrometry for urinary metabolic profiling. *Analytical Chemistry*, 84(5):2424–2432, Mar. 2012.
- [45] H. P. Benton, D. M. Wong, S. A. Trauger, and G. Siuzdak. XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization. *Analytical Chemistry*, 80(16):6382–6389, 2008.

- [46] J. H. Beynon. The use of the mass spectrometer for the identification of organic compounds. *Microchimica Acta*, 44(1):437–453, 1956.
- [47] S. Bhaskaran, J. A. Butler, S. Becerra, V. Fassio, M. Girotti, and S. L. Rea. Breaking *Caenorhabditis elegans* the easy way using the Balch homogenizer: An old tool for a new application. *Analytical Biochemistry*, 413(2):123–132, June 2011.
- [48] B. Biais, J. W. Allwood, C. Deborde, Y. Xu, M. Maucourt, B. Beauvoit, W. B. Dunn, D. Jacob, R. Goodacre, D. Rolin, and A. Moing. H-1 NMR, GC-ESI-TOFMS, and Data Set Correlation for Fruit Metabolomics: Application to Spatial Metabolite Analysis in Melon. *Analytical Chemistry*, 81(8):2884–2894, 2009.
- [49] A. T. Blades, M. G. Ikonomou, and P. Kebarle. Mechanism of electrospray mass spectrometry. Electrospray as an electrolysis cell. *Analytical Chemistry*, 63(19):2109–2114, Oct. 1991.
- [50] B. J. Blaise, J. Giacomotto, B. Elena, M.-E. Dumas, P. Toulhoat, L. Ségalat, and L. Emsley. Metabotyping of *Caenorhabditis elegans* reveals latent phenotypes. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 104(50):19808–19812, Dec. 2007.
- [51] B. J. Blaise, J. Giacomotto, M. N. Triba, P. Toulhoat, M. Piotto, L. Emsley, L. Ségalat, M.-E. Dumas, and B. Elena. Metabolic profiling strategy of *Caenorhabditis elegans* by whole-organism nuclear magnetic resonance. *Journal of Proteome Research*, 8(5):2542–2550, May 2009.
- [52] E. G. Bligh and W. J. Dyer. A rapid method of total lipid extraction and purification. *Canadian Journal of Biochemistry and Physiology*, 37(8):911–917, Aug. 1959.
- [53] M. Blüher, B. B. Kahn, and C. R. Kahn. Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science (New York, NY)*, 299(5606):572–574, Jan. 2003.
- [54] L. M. Blumberg and M. S. Klee. Method Translation and Retention Time Locking in Partition GC. *Analytical Chemistry*, 70(18):3828–3839, Sept. 1998.
- [55] G. Bodenhausen and D. Ruben. Natural Abundance Nitrogen-15 NMR by Enhanced Nuclear Spectroscopy. *Chemical Physics Letters*, 69(1):185–189, 1980.
- [56] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant. PubChem: integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry*, 4:217–241, 2008.
- [57] J. H. F. Bothwell and J. L. Griffin. An introduction to biological nuclear magnetic resonance spectroscopy. *Biological Reviews*, 86(2):493–510, Oct. 2010.

- [58] B. P. Bowen, C. R. Fischer, R. Baran, J. F. Banfield, and T. Northen. Improved genome annotation through untargeted detection of pathway-specific metabolites. *BMC Genomics*, 12 Suppl 1:S6, 2011.
- [59] D. Breitenstein, C. Rommel, and R. Möllers. Chemische Analyse tierischer Zellen und ihrer intrazellulären Kompartimente durch 3D-Massenspektrometrie. *Angewandte Chemie International Edition in English*, 119(28):5427–5431, 2007.
- [60] S. Brenner. The genetics of *Caenorhabditis elegans*. *Genetics*, 1974.
- [61] D. I. Broadhurst and D. B. Kell. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Chromatographia*, 2(4):171–196, Nov. 2006.
- [62] L. Brodsky, A. Moussaieff, N. Shahaf, A. Aharoni, and I. Rogachev. Evaluation of Peak Picking Quality in LC-MS Metabolomics Data. *Analytical Chemistry*, 82(22):9177–9187, Nov. 2010.
- [63] C. D. Broeckling, I. R. Reddy, A. L. Duran, X. Zhao, and L. W. Sumner. MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. *Analytical Chemistry*, 78(13):4334–4341, July 2006.
- [64] M. Brown, D. C. Wedge, R. Goodacre, D. B. Kell, P. N. Baker, L. C. Kenny, M. A. Mamas, L. Neyses, and W. B. Dunn. Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics (Oxford, England)*, 27(8):1108–1112, Apr. 2011.
- [65] A. P. Bruins. Mass spectrometry with ion sources operating at atmospheric pressure. *Mass Spectrometry Reviews*, 10(1):53–77, Jan. 1991.
- [66] K. Brys, N. Castelein, F. Matthijssens, J. R. Vanfleteren, and B. P. Braeckman. Disruption of insulin signalling preserves bioenergetic competence of mitochondria in ageing *Caenorhabditis elegans*. *BMC Biology*, 8(1):91, 2010.
- [67] J. M. Buescher, D. Czernik, J. C. Ewald, U. Sauer, and N. Zamboni. Cross-Platform Comparison of Methods for Quantitative Metabolomics of Primary Metabolism. *Analytical Chemistry*, 81(6):2135–2143, 2009.
- [68] J. M. Buescher, S. Moco, U. Sauer, and N. Zamboni. Ultrahigh Performance Liquid Chromatography-Tandem Mass Spectrometry Method for Fast and Robust Quantification of Anionic and Aromatic Metabolites. *Analytical Chemistry*, 82(11):4403–4412, 2010.
- [69] C. Bueschl, B. Kluger, F. Berthiller, G. Lirk, S. Winkler, R. Krska, and R. Schuhmacher. MetExtract: A new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research. *Bioinformatics (Oxford, England)*, Jan. 2012.

- [70] M. G. Bulmer. *Principles of statistics*. Dover Books on Mathematics Series. Dover Publications, Incorporated, 1979.
- [71] J. G. Bundy, J. K. Sidhu, F. Rana, D. J. Spurgeon, C. Svendsen, J. F. Wren, S. R. Sturzenbaum, A. J. Morgan, and P. Kille. 'Systems toxicology' approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm *Lumbricus rubellus*. *BMC Biology*, 6:25, 2008.
- [72] A. M. Burnell, K. Houthoofd, K. O'Hanlon, and J. R. Vanfleteren. Alternate metabolism during the dauer stage of the nematode *Caenorhabditis elegans*. *Experimental Gerontology*, 40(11):850–856, Nov. 2005.
- [73] C. Burnett, S. Valentini, F. Cabreiro, M. Goss, M. Somogyvári, M. D. Piper, M. Hoddinott, G. L. Sutphin, V. Leko, J. J. McElwee, R. P. Vazquez-Manrique, A.-M. Orfila, D. Ackerman, C. Au, G. Vinti, M. Riesen, K. Howard, C. Neri, A. Bedalov, M. Kaerberlein, C. Söti, L. Partridge, and D. Gems. Absence of effects of Sir2 overexpression on lifespan in *C. elegans* and *Drosophila*. *Nature*, 477(7365):482–485, Sept. 2011.
- [74] R. A. Butcher, M. Fujita, F. C. Schroeder, and J. Clardy. Small-molecule pheromones that control dauer development in *Caenorhabditis elegans*. *Nature Chemical Biology*, 3(7):420–422, July 2007.
- [75] J. A. Butler, R. J. Mishur, S. Bhaskaran, and S. L. Rea. A metabolic signature for long life in the *Caenorhabditis elegans* Mit mutants. *Aging Cell*, 12(1):130–138, Feb. 2013.
- [76] J. A. Butler, R. J. Mishur, A. F. Bokov, K. W. Hakala, S. T. Weintraub, and S. L. Rea. Profiling the anaerobic response of *C. elegans* using GC-MS. *PloS one*, 7(9):e46140, 2012.
- [77] J. A. Butler, N. Ventura, T. E. Johnson, and S. L. Rea. Long-lived mitochondrial (Mit) mutants of *Caenorhabditis elegans* utilize a novel metabolism. *The FASEB Journal*, 24:4977–4988, Aug. 2010.
- [78] L. BYERLY, R. C. Cassada, and R. L. Russell. The life cycle of the nematode *Caenorhabditis elegans*. *Developmental Biology*, 51(1):23–33, July 1976.
- [79] F. Cabreiro, C. Au, K.-Y. Leung, N. Vergara-Irigaray, H. M. Cochemé, T. Noori, D. Weinkove, E. Schuster, N. D. E. Greene, and D. Gems. Metformin retards aging in *C. elegans* by altering microbial folate and methionine metabolism. *Cell*, 153(1):228–239, Mar. 2013.
- [80] C. M. Cahill, G. Tzivion, N. Nasrin, S. Ogg, J. Dore, G. Ruvkun, and M. Alexander-Bridges. Phosphatidylinositol 3-Kinase Signaling Inhibits DAF-16 DNA Binding and Function via 14-3-3-dependent and 14-3-3-independent Pathways. *Journal Of Biological Chemistry*, 276(16):13402–13410, Dec. 2000.

- [81] A. E. Cameron and D. F. Eggers. An Ion“Velocitron”. *The Review of Scientific Instruments*, 19(9):605–607, 1948.
- [82] A. J. Carroll, M. R. Badger, and A. H. Millar. The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics*, 11:376, 2010.
- [83] X. Casadevall i Solvas, F. M. Geier, A. M. Leroi, J. G. Bundy, J. B. Edel, and A. J. Demello. High-throughput age synchronisation of *Caenorhabditis elegans*. *Chemical Communications*, 47(35):9801–9803, Sept. 2011.
- [84] C. Castro, J. Krumsiek, N. J. Lehrbach, S. A. Murfitt, E. A. Miska, and J. L. Griffin. A study of *Caenorhabditis elegans* DAF-2 mutants by metabolomics and differential correlation networks. *Molecular Biosystems*, 2013.
- [85] C. Castro, F. Sar, W. R. Shaw, M. Mishima, E. A. Miska, and J. L. Griffin. A metabolomic strategy defines the regulation of lipid content and global metabolism by ÎŦ9 desaturases in *Caenorhabditis elegans*. *BMC Genomics*, 13:36, 2012.
- [86] T. Cecchi. Ion pairing chromatography. *Critical Reviews in Analytical Chemistry*, 38(3):161–213, 2008.
- [87] N. B. Cech and C. G. Enke. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews*, 20(6):362–387, 2002.
- [88] *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science (New York, NY)*, 282(5396):2012–2018, Dec. 1998.
- [89] M. Chadeau-Hyam, T. M. D. Ebbels, I. J. Brown, Q. Chan, J. Stamler, C. C. Huang, M. L. Daviglius, H. Ueshima, L. Zhao, E. Holmes, J. K. Nicholson, P. Elliott, and M. De Iorio. Metabolic Profiling and the Metabolome-Wide Association Study: Significance Level For Biomarker Identification. *Analytical Chemistry*, 9(9):4620–4627, Sept. 2010.
- [90] A. Chalkiadaki and L. Guarente. Sirtuins mediate mammalian metabolic responses to nutrient availability. *Nature Reviews Endocrinology*, 8(5):287–296, May 2012.
- [91] H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber, and S. Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–547, May 2008.
- [92] D. Chen, E. L. Thomas, and P. Kapahi. HIF-1 modulates dietary restriction-mediated lifespan extension via IRE-1 in *Caenorhabditis elegans*. *PLoS genetics*, 5(5):e1000486, May 2009.

- [93] E. Chikayama, Y. Sekiyama, M. Okamoto, Y. Nakanishi, Y. Tsuboi, K. Akiyama, K. Saito, K. Shinozaki, and J. Kikuchi. Statistical indices for simultaneous large-scale metabolite detections for a single NMR spectrum. *Analytical Chemistry*, 82(5):1653–1658, Mar. 2010.
- [94] E. Chikayama, M. Suto, T. Nishihara, K. Shinozaki, T. Hirayama, and J. Kikuchi. Systematic NMR analysis of stable isotope labeled metabolite mixtures in plant and animal systems: coarse grained views of metabolic pathways. *PloS one*, 3(11):e3805, 2008.
- [95] S. Choe, S. H. Woo, D. W. Kim, Y. Park, H. Choi, B. Y. Hwang, D. Lee, and S. Kim. Development of a target component extraction method from GC-MS data with an in-house program for metabolite profiling. *Analytical Biochemistry*, 426(2):94–102, July 2012.
- [96] C. Choi, S. K. Ganji, R. J. DeBerardinis, K. J. Hatanpaa, D. Rakheja, Z. Kovacs, X.-L. Yang, T. Mashimo, J. M. Raisanen, I. Marin-Valencia, J. M. Pascual, C. J. Madden, B. E. Mickey, C. R. Malloy, R. M. Bachoo, and E. A. Maher. 2-hydroxyglutarate detection by magnetic resonance spectroscopy in IDH-mutated patients with gliomas. *Nature Medicine*, 18(4):624–629, Apr. 2012.
- [97] K. Christensen, G. Doblhammer, R. Rau, and J. W. Vaupel. Ageing populations: the challenges ahead. *Lancet*, 374(9696):1196–1208, Oct. 2009.
- [98] U. Christians, J. Klawitter, A. Hornberger, and J. Klawitter. How unbiased is non-targeted metabolomics and is targeted pathway screening the solution? *Current Pharmaceutical Biotechnology*, 12(7):1053–1066, July 2011.
- [99] C.-Y. C. L.-Y. C. C.-S. L. W.-c. L. Chun-Hung Lai. Identification of Novel Human Genes Evolutionarily Conserved in *Caenorhabditis elegans* by Comparative Proteomics. *Genome Research*, 10(5):703, May 2000.
- [100] K. Chung, M. M. Crane, and H. Lu. Automated on-chip rapid microscopy, phenotyping and sorting of *C. elegans*. *Nature Methods*, 5(7):637–643, July 2008.
- [101] M. Churchwell, N. Twaddle, L. Meeker, and D. Doerge. Improving LC-MS sensitivity through increases in chromatographic performance: Comparisons of UPLC-ES/MS/MS to HPLC-ES/MS/MS. *Journal Of Chromatography B-Analytical Technologies In The Biomedical And Life Sciences*, 825(2):134–143, 2005.
- [102] D. J. Clancy, D. Gems, L. G. Harshman, S. Oldham, H. Stocker, E. Hafen, S. J. Leevers, and L. Partridge. Extension of life-span by loss of CHICO, a *Drosophila* insulin receptor substrate protein. *Science (New York, NY)*, 292(5514):104–106, Apr. 2001.

- [103] T. D. W. Claridge. *High-resolution Nuclear Magnetic Resonance Techniques in Organic Chemistry*. Tetrahedron Organic Chemistry Series. Elsevier Science & Technology Books, 1999.
- [104] J. Clausell-Tormos, D. Lieber, J.-C. Baret, A. El-Harrak, O. J. Miller, L. Frenz, J. Blouwolff, K. J. Humphry, S. Köster, H. Duan, C. Holtze, D. A. Weitz, A. D. Griffiths, and C. A. Merten. Droplet-based microfluidic platforms for the encapsulation and screening of Mammalian cells and multicellular organisms. *Chemistry & Biology*, 15(5):427–437, May 2008.
- [105] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.
- [106] O. Cloarec, M.-E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes, and J. Nicholson. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^1H NMR data sets. *Analytical Chemistry*, 77(5):1282–1289, Mar. 2005.
- [107] J. J. Collins, K. Evason, and K. Kornfeld. Pharmacology of delayed aging and extended lifespan of *Caenorhabditis elegans*. *Experimental Gerontology*, 41(10):1032–1039, Oct. 2006.
- [108] J. J. Collins, C. Huang, S. Hughes, and K. Kornfeld. The measurement and analysis of age-related changes in *Caenorhabditis elegans*. *WormBook*, pages 1–21, 2008.
- [109] M. B. Comisarow and A. G. Marshall. Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters*, 25(2):282–283, 1974.
- [110] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*, 19(90):297–301, 1965.
- [111] O. Corcoran and M. Spraul. ScienceDirect.com - Drug Discovery Today - LC–NMR–MS in drug discovery. *Drug Discovery Today*, 2003.
- [112] D. Cornett, M. Reyzer, P. Chaurand, and R. Caprioli. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature Methods*, 4(10), 2007.
- [113] D. B. COWIE, E. T. BOLTON, and M. K. SANDS. Sulfur metabolism in *Escherichia coli*. I. Sulfate metabolism of normal and mutant cells. *Journal Of Bacteriology*, 60(3):233–248, Sept. 1950.
- [114] G. B. Cox, M. Kusch, and R. S. Edgar. Cuticle of *Caenorhabditis elegans* - Its Isolation and Partial Characterization. *Journal Of Cell Biology*, 90(1):7–17, 1981.
- [115] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–2267, Apr. 2006.

- [116] M. M. Crane, K. Chung, and H. Lu. Computer-enhanced high-throughput genetic screens of *C. elegans* in a microfluidic system. *Lab On A Chip*, 9(1):38–40, Jan. 2009.
- [117] D. J. Creek, A. Jankevics, R. Breitling, D. G. Watson, M. P. Barrett, and K. E. V. Burgess. Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: improved metabolite identification by retention time prediction. *Analytical Chemistry*, 83(22):8703–8710, Nov. 2011.
- [118] B. Crews, W. R. Wikoff, G. J. Patti, H.-K. Woo, E. Kalisiak, J. Heideker, and G. Siuzdak. Variability analysis of human plasma and cerebral spinal fluid reveals statistical significance of changes in mass spectrometry-based metabolomics data. *Analytical Chemistry*, 81(20):8538–8544, Oct. 2009.
- [119] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug. 1970.
- [120] X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4):210, 2003.
- [121] S. P. Curran and G. Ruvkun. Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS genetics*, 3(4):e56, Apr. 2007.
- [122] C. E. Dalgliesh, E. C. Horning, M. G. Horning, K. L. Knox, and K. Yarger. A gas-liquid-chromatographic procedure for separating a wide range of metabolites occurring in urine or tissue extracts. *The Biochemical journal*, 101(3):792–810, Dec. 1966.
- [123] R. Danielsson, D. Bylund, and K. Markides. Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography-mass spectrometry. *Analytica Chimica Acta*, 454(2):167–184, 2002.
- [124] G. D’Antona, M. Ragni, A. Cardile, L. Tedesco, M. Dossena, F. Bruttini, F. Caliaro, G. Corsetti, R. Bottinelli, M. O. Carruba, A. Valerio, and E. Nisoli. Branched-chain amino acid supplementation promotes survival and supports cardiac and skeletal muscle mitochondrial biogenesis in middle-aged mice. *Cell Metabolism*, 12(4):362–372, Oct. 2010.
- [125] S. Davies, A. Leroi, and J. G. Bundy. Fluorodeoxyuridine affects the identification of metabolic responses to DAF-2 status in *Caenorhabditis elegans*. *Mechanisms Of Ageing And Development*, 2011.
- [126] S. K. Davies. *A Metabolomic Approach to Assessing Life-History Traits in Caenorhabditis elegans*. PhD thesis, Imperial College London, London, July 2011.
- [127] B. O. Davis, G. L. Anderson, and D. B. Dusenbery. Total Luminescence Spectroscopy of Fluorescence Changes during aging in *Caenorhabditis elegans*. *Biochemistry*, 21(17):4089–4095, 1982.

- [128] J. M. De Guzman, G. Ku, R. Fahey, Y.-H. Youm, I. Kass, D. K. Ingram, V. D. Dixit, and I. Kheterpal. Chronic caloric restriction partially protects against age-related alteration in serum metabolome. *Age*, June 2012.
- [129] E. de Hoffmann. Tandem mass spectrometry: a primer. *Journal of Mass Spectrometry*, 31(2):129–137, 1996.
- [130] E. de Hoffmann and V. Stroobant. *Mass spectrometry: Principles and applications*. J. Wiley & Sons, New York, 3 edition, 1997.
- [131] F. A. de Jong and C. Beecher. Addressing the current bottlenecks of metabolomics: Isotopic Ratio Outlier AnalysisTM, an isotopic-labeling technique for accurate biochemical profiling. *Bioanalysis*, 4(18):2303–2314, Sept. 2012.
- [132] J. R. De Laeter, J. K. Böhlke, P. De Bièvre, H. Hidaka, H. S. Peiser, K. Rosman, and P. Taylor. Atomic weights of the elements: Review 2000. *Pure Appl. Chem*, 75(6):683–800, 2003.
- [133] G. J. Dear, R. S. Plumb, B. C. Sweatman, P. S. Parry, A. D. Roberts, J. C. Lindon, J. K. Nicholson, and I. M. Ismail. Use of directly coupled ion-exchange liquid chromatography-mass spectrometry and liquid chromatography-nuclear magnetic resonance spectroscopy as a strategy for polar metabolite identification. *Journal of Chromatography. B*, 748(1):295–309, Oct. 2000.
- [134] A. J. Dempster. A new method of positive ray analysis. *Physical Review*, 11(4):316–325, 1918.
- [135] K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78, 2007.
- [136] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, 78:4281–4290, 2006.
- [137] A. Dillin, A.-L. Hsu, N. Arantes-Oliveira, J. Lehrer-Graiwer, H. Hsin, A. G. Fraser, R. S. Kamath, J. Ahringer, and C. Kenyon. Rates of behavior and aging specified by mitochondrial function during development. *Science (New York, NY)*, 298(5602):2398–2401, Dec. 2002.
- [138] C. Dong, K. Beis, M.-F. Giraud, W. Blankenfeldt, S. Allard, L. L. Major, I. D. Kerr, C. Whitfield, and J. H. Naismith. A structural perspective on the enzymes that convert dTDP-d-glucose into dTDP-l-rhamnose. *Biochemical Society transactions*, 31(Pt 3):532–536, June 2003.
- [139] R. Doonan, J. J. McElwee, F. Matthijssens, G. A. Walker, K. Houthoofd, P. Back, A. Matscheski, J. R. Vanfleteren, and D. Gems. Against the oxidative damage theory

- of aging: superoxide dismutases protect against oxidative stress but have little or no effect on life span in *Caenorhabditis elegans*. *Genes & Development*, 22(23):3236–3241, Dec. 2008.
- [140] R. Dromey, M. Stefik, and T. Rindfleisch. Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data. *Analytical Chemistry*, 48(9):1368–1375, 1976.
- [141] W. B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J. D. Knowles, A. Halsall, J. N. Haselden, A. W. Nicholls, I. D. Wilson, D. B. Kell, R. Goodacre, and Human Serum Metabolome (HUSERMET) Consortium. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, 6(7):1060–1083, June 2011.
- [142] W. B. Dunn, D. Broadhurst, M. Brown, P. N. Baker, C. W. G. Redman, L. C. Kenny, and D. B. Kell. Metabolic profiling of serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system. *International Journal of Pharmaceutics*, 871(2):288–298, Aug. 2008.
- [143] W. B. Dunn, D. I. Broadhurst, H. J. Atherton, R. Goodacre, and J. L. Griffin. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*, 40(1):387–426, Jan. 2011.
- [144] A. L. Duran, J. Yang, L. Wang, and L. W. Sumner. Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics (Oxford, England)*, 19(17):2283–2293, Nov. 2003.
- [145] B. Durbin and D. M. Rocke. Estimation of transformation parameters for microarray data. *Bioinformatics (Oxford, England)*, 19(11):1360–1367, July 2003.
- [146] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S105–10, 2002.
- [147] J. Durieux, S. Wolff, and A. Dillin. The Cell-Non-Autonomous Nature of Electron Transport Chain-Mediated Longevity. *Cell*, 144(1):79–91, Jan. 2011.
- [148] B. Dutta, H. Kanani, J. Quackenbush, and M. Klapa. Time-series integrated “omic” analyses to elucidate short-term stress-induced responses in plant liquid cultures. *Biotechnology And Bioengineering*, 102(1):264–279, 2009.
- [149] R. S. Edgar and W. B. Wood. The nematode *Caenorhabditis elegans*: a new organism for intensive biological study. *Science (New York, NY)*, 198(4323):1285–1286, Dec. 1977.

- [150] H. M. Ellis and H. R. Horvitz. Genetic control of programmed cell death in the nematode *C. elegans*. *Cell*, 44(6):817–829, Mar. 1986.
- [151] E. V. Entchev and T. V. Kurzchalia. Requirement of sterols in the life cycle of the nematode *Caenorhabditis elegans*. *Seminars in Cell & Developmental Biology*, 16(2):175–182, Apr. 2005.
- [152] A. C. Epstein, J. M. Gleadle, L. A. McNeill, K. S. Hewitson, J. O’Rourke, D. R. Mole, M. Mukherji, E. Metzen, M. I. Wilson, A. Dhanda, Y. M. Tian, N. Masson, D. L. Hamilton, P. Jaakkola, R. Barstead, J. Hodgkin, P. H. Maxwell, C. W. Pugh, C. J. Schofield, and P. J. Ratcliffe. *C. elegans* EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation. *Cell*, 107(1):43–54, Oct. 2001.
- [153] R. R. Ernst. Application of Fourier Transform Spectroscopy to Magnetic Resonance. *The Review of Scientific Instruments*, 37(1):93, 1966.
- [154] N. Etxebarria, O. Zuloaga, M. Olivares, L. J. Bartolomé, and P. Navarro. Retention-time locked methods in gas chromatography. *Journal Of Chromatography A*, 1216(10):1624–1629, Mar. 2009.
- [155] K. Evason, J. Collins, C. Huang, S. Hughes, and K. Kornfeld. Valproic acid extends *Caenorhabditis elegans* lifespan. *Aging Cell*, 2008.
- [156] E. S. A. K. S. J. A. W. Ewa Szymańska. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, 8(Suppl 1):3, June 2012.
- [157] V. Exarchou, M. Krucker, T. A. van Beek, J. Vervoort, I. P. Gerothanassis, and K. Albert. LC-NMR coupling technology: recent advancements and applications in natural products analysis. *Magnetic Resonance In Chemistry*, 43(9):681–687, Sept. 2005.
- [158] P. Fagone and S. Jackowski. Phosphatidylcholine and the CDP-choline cycle. *Biochimica et Biophysica Acta*, 1831(3):523–532, Mar. 2013.
- [159] M. J. Falk, Z. Zhang, J. R. Rosenjack, E. Daikhin, I. Nissim, M. M. Sedensky, M. Yudkoff, and P. G. Morgan. Metabolic pathway profiling of mitochondrial respiratory chain mutants in *C. elegans*. *Molecular Genetics And Metabolism*, 93(4):388–397, Apr. 2008.
- [160] W. Fan. Metabolite profiling by one- and two-dimensional NMR analysis of complex mixtures. *Progress In Nuclear Magnetic Resonance Spectroscopy*, 28:161–219, 1996.
- [161] X. Fan, J. Zhang, M. Theves, C. Strauch, I. Nemet, X. Liu, J. Quian, F. Giblin, and V. Monnier. Mechanism of lysine oxidation in human lens crystallins during aging and in diabetes. *Journal Of Biological Chemistry*, 2009.

- [162] M. K. Fasseas, C. Fasseas, K. C. Mountzouris, and P. Syntichaki. Effects of *Lactobacillus salivarius*, *Lactobacillus reuteri*, and *Pediococcus acidilactici* on the nematode *Caenorhabditis elegans* include possible antitumor activity. *Applied Microbiology and Biotechnology*, 97(5):2109–2118, Aug. 2012.
- [163] J.-L. Faulon. Stochastic Generator of Chemical Structure. 1. Application to the Structure Elucidation of Large Molecules. *Journal of Chemical Information and Modeling*, 34(5):1204–1218, Sept. 1994.
- [164] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- [165] FDA. Guidance for Industry - Bioanalytical Method Validation. Technical report, Centre for Drug Valuation and Research, May 2001.
- [166] S. Felkai, J. Ewbank, J. Lemieux, J. Labbe, G. Brown, and S. Hekimi. CLK-1 controls respiration, behavior and aging in the nematode *Caenorhabditis elegans*. *Embo Journal*, 18(7):1783–1792, 1999.
- [167] S. Fendt, M. Rühl, and U. Sauer. ¹³C-based metabolic flux analysis. *Nature Protocols*, 2009.
- [168] S.-M. Fendt, J. M. Buescher, F. Rudroff, P. Picotti, N. Zamboni, and U. Sauer. Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Molecular Systems Biology*, 6:356, Apr. 2010.
- [169] J. Feng, F. Bussière, and S. Hekimi. Mitochondrial electron transport is a key determinant of life span in *Caenorhabditis elegans*. *Developmental Cell*, 1(5):633–644, Nov. 2001.
- [170] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science (New York, NY)*, 246(4926):64–71, Oct. 1989.
- [171] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey, and L. Willmitzer. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, 18(11):1157–1161, Nov. 2000.
- [172] O. Fiehn, G. Wohlgemuth, and M. Scholz. *Setup and Annotation of Metabolomic Experiments by Integrating Biological and Mass Spectrometric Metadata*, volume 3615 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
- [173] T. M. Fink and Y. Mao. Designing tie knots by random walks. *Nature*, 398(6722):31–32, 1999.

- [174] T. Finkel and N. Holbrook. Oxidants, oxidative stress and the biology of ageing. *Nature*, 408:239–247, 2000.
- [175] E. Fischer and U. Sauer. FiatFlux – a software for metabolic flux analysis from 13 C-glucose experiments. *BMC Bioinformatics*, 2005.
- [176] R. A. Fisher. *Statistical Methods for Research Workers*. Cosmo study guides. Cosmo Publications, 1925.
- [177] J. S. Fletcher, N. P. Lockyer, S. Vaidyanathan, and J. C. Vickerman. TOF-SIMS 3D biomolecular imaging of *Xenopus laevis* oocytes using buckminsterfullerene (C60) primary ions. *Analytical Chemistry*, 79(6):2199–2206, Mar. 2007.
- [178] J. S. Fletcher and J. C. Vickerman. A new SIMS paradigm for 2D and 3D molecular imaging of bio-systems. *Analytical and Bioanalytical Chemistry*, 396(1):85–104, 2010.
- [179] L. Fontana. Long-term calorie restriction is highly effective in reducing the risk for atherosclerosis in humans. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 101(17):6659–6663, Apr. 2004.
- [180] M. Fredriksson, P. Petersson, M. Jörntén-Karlsson, B.-O. Axelsson, and D. Bylund. An objective comparison of pre-processing methods for enhancement of liquid chromatography-mass spectrometry data. *Journal Of Chromatography A*, 1172(2):135–150, Nov. 2007.
- [181] J. Freedman, L. Slice, D. Dixon, A. Fire, and C. Rubin. The novel metallothionein genes of *Caenorhabditis elegans*. Structural organization and inducible, cell-specific expression. *Journal Of Biological Chemistry*, 268(4), 1993.
- [182] D. B. Friedman and T. E. Johnson. A mutation in the age-1 gene in *Caenorhabditis elegans* lengthens life and reduces hermaphrodite fertility. *Genetics*, 118(1):75–86, Jan. 1988.
- [183] S. Fuchs, J. G. Bundy, S. K. Davies, J. M. Viney, J. S. Swire, and A. M. Leroi. A metabolic signature of long life in *Caenorhabditis elegans*. *BMC Biology*, 8:14, 2010.
- [184] K. H. Gardner and L. E. Kay. The use of 2H, 13C, 15N Multidimensional NMR To Study the Structure and Dynamics of Proteins. *Annu. Rev. Biophys. Biomol. Struct.*, 27:357–406, Jan. 1998.
- [185] F. M. Geier, S. Fearn, J. G. Bundy, and D. S. McPhail. ToF-SIMS analysis of biomolecules in the model organism *Caenorhabditis elegans*. *Surface and Interface Analysis*, 45(1):234–236, 2012.
- [186] F. M. Geier, S. Fuchs, G. Valbuena, A. M. Leroi, and J. G. Bundy. Profiling the Metabolic Signature of Senescence. In *Cellular Senescence*, pages 355–371. Methods in Molecular Biology, 2013.

- [187] F. M. Geier, E. J. Want, A. M. Leroi, and J. G. Bundy. Cross-Platform Comparison of *Caenorhabditis elegans* Tissue Extraction Strategies for Comprehensive Metabolome Coverage. *Analytical Chemistry*, 83(10):3730–3736, May 2011.
- [188] S. Geisser and W. O. Johnson. *Modes of Parametric Statistical Inference*. Wiley Series in Probability and Statistics. Wiley, 2006.
- [189] D. Gems and L. Partridge. Stress-Response Hormesis and Aging: “That which Does Not Kill Us Makes Us Stronger”. *Cell Metabolism*, 7(3):200–203, Mar. 2008.
- [190] D. Gems and L. Partridge. Genetics of Longevity in Model Organisms: Debates and Paradigm Shifts. *Annual Reviews in Physiology*, 75:621–644, Mar. 2013.
- [191] G. S. Gerhard. Small laboratory fish as models for aging research. *Ageing Research Reviews*, 6(1):64–72, May 2007.
- [192] B. Gerisch, V. Rottiers, D. Li, D. L. Motola, C. L. Cummins, H. Lehrach, D. J. Mangelsdorf, and A. Antebi. A bile acid-like steroid modulates *Caenorhabditis elegans* lifespan through nuclear receptor signaling. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 104(12):5014–5019, Mar. 2007.
- [193] M. Gerlich and S. Neumann. MetFusion: integration of compound identification strategies. *Journal of Mass Spectrometry*, 48(3):291–298, Mar. 2013.
- [194] P. Giavalisco, J. Hummel, K. Köhl, and L. Willmitzer. Elemental formula annotation of polar and lipophilic metabolites using (13) C, (15) N and (34) S isotope labelling, in combination with high-resolution mass spectrometry. *The Plant journal*, 68(2):364–376, Oct. 2011.
- [195] P. Giavalisco, J. Hummel, J. Lisec, A. C. Inostroza, G. Catchpole, and L. Willmitzer. High-resolution direct infusion-based mass spectrometry in combination with whole ¹³C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. *Analytical Chemistry*, 80(24):9417–9425, Dec. 2008.
- [196] P. Giavalisco, J. Hummel, B. Seiwert, and L. Willmitzer. C-13 Isotope-Labeled Metabolomes Allowing for Improved Compound Annotation and Relative Quantification in Liquid Chromatography-Mass Spectrometry-based Metabolomic Research. *Analytical Chemistry*, 81(15):6546–6551, 2009.
- [197] H. G. Gika, E. Macpherson, G. A. Theodoridis, and I. D. Wilson. Evaluation of the repeatability of ultra-performance liquid chromatography-TOF-MS for global metabolic profiling of human urine samples. *Journal Of Chromatography B-Analytical Technologies In The Biomedical And Life Sciences*, 871(2):299–305, Aug. 2008.
- [198] H. G. Gika, G. A. Theodoridis, M. Earll, R. W. Snyder, S. J. Sumner, and I. D. Wilson. Does the Mass Spectrometer Define the Marker? A Comparison of Global

- Metabolite Profiling Data Generated Simultaneously via UPLC-MS on Two Different Mass Spectrometers. *Analytical Chemistry*, 82(19):8226–8234, Sept. 2010.
- [199] H. G. Gika, G. A. Theodoridis, J. E. Wingate, and I. D. Wilson. Within-day reproducibility of an HPLC-MS-based method for metabonomic analysis: application to human urine. *Journal of Proteome Research*, 6(8):3291–3303, Aug. 2007.
- [200] V. N. Gladyshev, M. Krause, X.-M. Xu, K. V. Korotkov, G. V. Kryukov, Q.-A. Sun, B. J. Lee, J. C. Wootton, and D. L. Hatfield. Selenocysteine-Containing Thioredoxin Reductase in *C. elegans*. *Biochemical and Biophysical Research Communications*, 259(2):244–249, June 1999.
- [201] M. Godejohann. Hydrophilic interaction chromatography coupled to nuclear magnetic resonance spectroscopy and mass spectroscopy—a new approach for the separation and identification of extremely polar analytes in bodyfluids. *Journal Of Chromatography A*, 1156(1-2):87–93, July 2007.
- [202] R. S. Gohlke. Time-of-flight mass spectrometry and gas-liquid partition chromatography. *Analytical Chemistry*, 31(4):535–541, 1959.
- [203] J. W. Golden and D. L. Riddle. The *Caenorhabditis elegans* dauer larva: developmental effects of pheromone, food, and temperature. *Developmental Biology*, 102(2):368–378, Apr. 1984.
- [204] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends In Biotechnology*, 22(5):245–252, May 2004.
- [205] H. E. Gottlieb, V. Kotlyar, and A. Nudelman. NMR Chemical Shifts of Common Laboratory Solvents as Trace Impurities. *The Journal of Organic Chemistry*, 62(21):7512–7515, Oct. 1997.
- [206] G. A. N. Gowda, F. Tayyari, T. Ye, Y. Suryani, S. Wei, N. Shanaiah, and D. Raftery. Quantitative Analysis of Blood Plasma Metabolites Using Isotope Enhanced NMR Methods. *Analytical Chemistry*, 82(21):8983–8990, Sept. 2010.
- [207] G. Graça, I. F. Duarte, B. J Goodfellow, I. M. Carreira, A. B. Couceiro, M. d. R. Domingues, M. Spraul, L.-H. Tseng, and A. M. Gil. Metabolite profiling of human amniotic fluid by hyphenated nuclear magnetic resonance spectroscopy. *Analytical Chemistry*, 80(15):6085–6092, Aug. 2008.
- [208] R. C. Grandison, M. D. W. Piper, and L. Partridge. Amino-acid imbalance explains extension of lifespan by dietary restriction in *Drosophila*. *Nature*, 462(7276):1061–1064, Dec. 2009.

- [209] F. M. Green, I. S. Gilmore, and M. P. Seah. TOF-SIMS: accurate mass scale calibration. *Journal Of The American Society For Mass Spectrometry*, 17(4):514–523, Apr. 2006.
- [210] E. L. Greer and A. Brunet. Different dietary restriction regimens extend lifespan by both independent and overlapping genetic pathways in *C. elegans*. *Aging Cell*, 8(2):113–127, Apr. 2009.
- [211] J. Griffin. Metabolic profiles to define the genome: can we hear the phenotypes? *Philosophical Transactions Of The Royal Society Of London Series B-Biological Sciences*, 359(1446):857–871, 2004.
- [212] L. Griffiths. Assay by nuclear magnetic resonance spectroscopy: quantification limits. *Analyst*, 123(5):1061–1068, 1998.
- [213] H. Gu, Z. Pan, B. Xi, B. E. Hainline, N. Shanaiah, V. Asiago, G. A. N. Gowda, and D. Raftery. ¹H NMR metabolomics study of age profiling in children. *Nmr In Biomedicine*, 22(8):826–833, Oct. 2009.
- [214] J. Gullberg, P. Jonsson, A. Nordstrom, M. Sjöström, and T. Moritz. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. *Analytical Biochemistry*, 331(2):283–295, Aug. 2004.
- [215] A. Gunnarsson, F. Kollmer, S. Sohn, F. Höök, and P. Sjövall. Spatial-resolution limits in mass spectrometry imaging of supported lipid bilayers and individual lipid vesicles. *Analytical Chemistry*, 82(6):2426–2433, Mar. 2010.
- [216] X. Guo, A. P. Bruins, and T. R. Covey. Characterization of typical chemical background interferences in atmospheric pressure ionization liquid chromatography-mass spectrometry. *Rapid Communications in Mass Spectrometry*, 20(20):3145–3150, 2006.
- [217] S. A. Gustavsson, J. Samskog, K. E. Markides, and B. Långström. Studies of signal suppression in liquid chromatography-electrospray ionization mass spectrometry using volatile ion-pairing reagents. *Journal Of Chromatography A*, 937(1-2):41–47, Dec. 2001.
- [218] D. Gutnick, J. M. Calvo, T. Klotowski, and B. N. Ames. Compounds which serve as the sole source of carbon or nitrogen for *Salmonella typhimurium* LT-2. *Journal Of Bacteriology*, 100(1):215–219, Oct. 1969.
- [219] J. Habbema, M. Eijkemans, G. Nargund, G. Beets, and H. Leridon. The effect of in vitro fertilization on birth rates in western countries. *Human Reproduction*, 6:1414–1419, 2009.

- [220] J. M. Halket, D. Waterman, A. M. Przyborowska, R. K. P. Patel, P. D. Fraser, and P. M. Bramley. Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *Journal of Experimental Botany*, 56(410):219–243, Jan. 2005.
- [221] E. A. Hallem and P. W. Sternberg. Acute carbon dioxide avoidance in *Caenorhabditis elegans*. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 105(23):8038–8043, June 2008.
- [222] M. Hamasaki, N. Furuta, A. Matsuda, A. Nezu, A. Yamamoto, N. Fujita, H. Oomori, T. Noda, T. Haraguchi, Y. Hiraoka, A. Amano, and T. Yoshimori. Autophagosomes form at ER–mitochondria contact sites. *Nature*, 495(7441):389–393, Mar. 2013.
- [223] A. J. Hamilton and D. C. Baulcombe. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science (New York, NY)*, 286(5441):950–952, Oct. 1999.
- [224] B. Hamilton, Y. Dong, M. Shindo, W. Liu, I. Odell, G. Ruvkun, and S. S. Lee. A systematic RNAi screen for longevity genes in *C. elegans*. *Genes & Development*, 19(13):1544–1555, July 2005.
- [225] M. Hansen, A. Chandra, L. L. Mitic, B. Onken, M. Driscoll, and C. Kenyon. A role for autophagy in the extension of lifespan by dietary restriction in *C. elegans*. *PLoS genetics*, 4(2):e24, Feb. 2008.
- [226] M. Hansen, T. Flatt, and H. Aguilaniu. Reproduction, Fat Metabolism, and Life Span: What Is the Connection? *Cell Metabolism*, 17(1):10–19, Jan. 2013.
- [227] M. Hansen, A.-L. Hsu, A. Dillin, and C. Kenyon. New genes tied to endocrine, metabolic, and dietary regulation of lifespan from a *Caenorhabditis elegans* genomic RNAi screen. *PLoS genetics*, 1(1):119–128, July 2005.
- [228] M. Hansen, S. Taubert, D. Crawford, N. Libina, S.-J. Lee, and C. Kenyon. Lifespan extension by conditions that inhibit translation in *Caenorhabditis elegans*. *Aging Cell*, 6(1):95–110, Feb. 2007.
- [229] J. Hao, W. Astle, M. De Iorio, and T. M. D. Ebbels. BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics (Oxford, England)*, 28(15):2088–2090, 2012.
- [230] D. HARMAN. Aging: a theory based on free radical and radiation chemistry. *Journal of Gerontology*, 11(3):298–300, July 1956.
- [231] D. C. Harris. *Quantitative Chemical Analysis, Sixth Edition*. W. H. Freeman, 2003.
- [232] E. S. Hars, H. Qi, A. G. Ryazanov, S. Jin, L. Cai, C. Hu, and L. F. Liu. Autophagy regulates ageing in *C. elegans*. *Autophagy*, 3(2):93–95, Mar. 2007.

- [233] R. M. A. Heeren, L. A. McDonnell, E. Amstalden, S. L. Luxembourg, A. F. M. Altelaar, and S. R. Piersma. Why don't biologists use SIMS? A critical evaluation of imaging MS. *Applied Surface Science*, 252(19):6827–6835, 2006.
- [234] A. D. Hegeman, C. F. Schulte, Q. Cui, I. A. Lewis, E. L. Huttlin, H. Eghbalnia, A. C. Harms, E. L. Ulrich, J. L. Markley, and M. R. Sussman. Stable isotope assisted assignment of elemental compositions for metabolomics. *Analytical Chemistry*, 79(18):6912–6921, Sept. 2007.
- [235] T. Heidler, K. Hartwig, and H. Daniel. *Caenorhabditis elegans* lifespan extension caused by treatment with an orally active ROS-generator is dependent on DAF-16 and SIR-2.1 - Springer. *Biogerontology*, 2010.
- [236] S. Hekimi, J. Lapointe, and Y. Wen. Taking a “good” look at free radicals in the aging process. *Trends in Cell Biology*, 21(10):569–576, Oct. 2011.
- [237] P. Hemstrom and K. Irgum. Hydrophilic interaction chromatography. *Journal Of Separation Science*, 29(12):1784–1821, Aug. 2006.
- [238] S. T. Henderson and T. E. Johnson. daf-16 integrates developmental and environmental inputs to mediate aging in the nematode *Caenorhabditis elegans*. *Current Biology*, 11(24):1975–1980, Dec. 2001.
- [239] R. K. Herman and T. L. Stiernagle. Status of the *Caenorhabditis* Genetics Center (CGC). *Worm Breeder's Gazette*, 14, 1997.
- [240] A. M. Herskind, M. McGue, I. A. Iachine, N. Holm, T. I. Sørensen, B. Harvald, and J. W. Vaupel. Untangling genetic influences on smoking, body mass index and longevity: a multivariate study of 2464 Danish twins followed for 28 years. *Human Genetics*, 98(4):467–475, Oct. 1996.
- [241] M. Hertweck, C. Göbel, and R. Baumeister. *C. elegans* SGK-1 is the critical component in the Akt/PKB kinase complex to control stress response and life span. *Developmental Cell*, 6(4):577–588, Apr. 2004.
- [242] S. L. Hillyard and J. B. German. Quantitative lipid analysis and life span of the fat-3 mutant of *Caenorhabditis elegans*. *Journal of Agricultural and Food Chemistry*, 57(8):3389–3396, Apr. 2009.
- [243] T. Hirose. Cyclic GMP-dependent protein kinase EGL-4 controls body size and lifespan in *C. elegans*. *Development*, 130(6):1089–1099, Mar. 2003.
- [244] J. Hodgkin. 9 Sexual Dimorphism and Sex Determination. *Cold Spring Harbor Monograph Archive*, 1988.

- [245] E. Holmes, R. L. Loo, J. Stamler, M. Bictash, I. K. S. Yap, Q. Chan, T. Ebbels, M. De Iorio, I. J. Brown, K. A. Veselkov, M. L. Daviglus, H. Kesteloot, H. Ueshima, L. Zhao, J. K. Nicholson, and P. Elliott. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*, 453(7193):396–400, Apr. 2008.
- [246] E. Holmes, I. D. Wilson, and J. K. Nicholson. Metabolic phenotyping in health and disease. *Cell*, 134(5):714–717, 2008.
- [247] M. Holzenberger, J. Dupont, B. Ducos, P. Leneuve, A. G  lo  n, P. C. Even, P. Cervera, and Y. Le Bouc. IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature*, 421(6919):182–187, Jan. 2003.
- [248] S. Honda, N. Ishii, K. Suzuki, and M. Matsuo. Oxygen-dependent perturbation of life span and aging rate in the nematode. *Journal of Gerontology*, 48(2):B57–61, Mar. 1993.
- [249] Y. Honda, M. Tanaka, and S. Honda. Trehalose extends longevity in the nematode *Caenorhabditis elegans*. *Aging Cell*, 4(9), 2010.
- [250] S. Honjoh, T. Yamamoto, M. Uno, and E. Nishida. Signalling through RHEB-1 mediates intermittent fasting-induced longevity in *C. elegans*. *Nature*, 457(7230):726–730, Feb. 2009.
- [251] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, July 2010.
- [252] P. J. Hore, J. J. A. Jones, and S. Wimperis. *NMR.:* Oxford Chemistry Primers Series. Oxford University Press, Incorporated, 2000.
- [253] R. S. Houk, V. A. Fassel, G. D. Flesch, H. J. Svec, A. L. Gray, and C. E. Taylor. Inductively coupled argon plasma as an ion source for mass spectrometric determination of trace elements. *Analytical Chemistry*, 52(14):2283–2289, 1980.
- [254] K. Houthoofd, B. P. Braeckman, I. Lenaerts, K. Brys, A. De Vreese, S. Van Eygen, and J. R. Vanfleteren. Ageing is reversed, and metabolism is reset to young levels in recovering dauer larvae of *C. elegans*. *Experimental Gerontology*, 37(8-9):1015–1021, Aug. 2002.
- [255] K. Houthoofd and J. R. Vanfleteren. Public and private mechanisms of life extension in *Caenorhabditis elegans*. *Molecular Genetics and Genomics*, 277(6):601–617, June 2007.

- [256] R. H. Houtkooper, C. Argmann, S. M. Houten, C. Cantó, E. H. Jeninga, P. A. Andreux, C. Thomas, R. Doenlen, K. Schoonjans, and J. Auwerx. The metabolic footprint of aging in mice. *Scientific Reports*, 1:134, 2011.
- [257] K. T. Howitz, K. J. Bitterman, H. Y. Cohen, D. W. Lamming, S. Lavu, J. G. Wood, R. E. Zipkin, P. Chung, A. Kisielewski, L.-L. Zhang, B. Scherer, and D. A. Sinclair. Small molecule activators of sirtuins extend *Saccharomyces cerevisiae* lifespan. *Nature*, 425(6954):191–196, Sept. 2003.
- [258] H. Hsin and C. Kenyon. Signals from the reproductive system regulate the lifespan of *C. elegans*. *Nature*, 399(6734):362–366, May 1999.
- [259] A. Hsu. Diophantine Approach to Isotopic Abundance Calculations. *Analytical Chemistry*, 56(8):1356–1361, 1984.
- [260] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks. The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, 40(4):430–443, 2005.
- [261] J. Huang and B. D. Lemire. Mutations in the *C. elegans* succinate dehydrogenase iron-sulfur subunit promote superoxide generation and premature aging. *Journal of Molecular Biology*, 387(3):559–569, Apr. 2009.
- [262] M.-Z. Huang, S.-C. Cheng, Y.-T. Cho, and J. Shiea. Ambient ionization mass spectrometry: a tutorial. *Analytica Chimica Acta*, 702(1):1–15, Sept. 2011.
- [263] K. Hughes and R. Reynolds. Evolutionary and mechanistic theories of aging. *Annual Review Of Entomology*, 50:421–445, 2005.
- [264] S. L. Hughes, J. G. Bundy, E. J. Want, P. Kille, and S. R. Stürzenbaum. The metabolomic responses of *Caenorhabditis elegans* to cadmium are largely independent of metallothionein status, but dominated by changes in cystathionine and phytochelatins. *Journal of Proteome Research*, 8(7):3512–3519, July 2009.
- [265] A. J. Hulbert. Longevity, lipids and *C. elegans*. *Aging*, 3(2):81–82, Feb. 2011.
- [266] S. E. Hulme, S. S. Shevkoplyas, A. P. McGuigan, J. Apfeld, W. Fontana, and G. M. Whitesides. Lifespan-on-a-chip: microfluidic chambers for performing lifelong observation of *C. elegans*. *Lab On A Chip*, 10(5):589–597, 2010.
- [267] S. E. Hulme, S. S. Shevkoplyas, and A. Samuel. Microfluidics: streamlining discovery in worm biology. *Nature Methods*, 5(7):589–590, July 2008.
- [268] A. B. Hwang and S.-J. Lee. Regulation of life span by mitochondrial respiration: the HIF-1 and ROS connection. *Aging*, 3(3):304–310, Mar. 2011.

- [269] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*, 2:343–372, 2001.
- [270] S. Imai, C. M. Armstrong, M. Kaerberlein, and L. Guarente. Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature*, 403(6771):795–800, Feb. 2000.
- [271] P. N. Inglis, G. Ou, M. R. Leroux, and J. M. Scholey. The sensory cilia of *Caenorhabditis elegans*. *WormBook*, 2006.
- [272] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct. 2004.
- [273] R. L. Isaacson, P. J. Simpson, M. Liu, E. Cota, X. Zhang, P. Freemont, and S. Matthews. A New Labeling Method for Methyl Transverse Relaxation-Optimized Spectroscopy NMR Spectra of Alanine Residues. *Journal of the American Chemical Society*, 129(50):15428–15429, Dec. 2007.
- [274] N. Ishii, M. Fujii, P. S. Hartman, M. Tsuda, K. Yasuda, N. Senoo-Matsuda, S. Yanase, D. Ayusawa, and K. Suzuki. A mutation in succinate dehydrogenase cytochrome b causes oxidative stress and ageing in nematodes. *Nature*, 394(6694):694–697, 1998.
- [275] H. J. Issaq, Q. N. Van, T. J. Waybright, G. M. Muschik, and T. D. Veenstra. Analytical and statistical approaches to metabolomics research. *Proteomics*, 32(13):2183–2199, June 2009.
- [276] Y. Izrayelit, J. Srinivasan, S. L. Campbell, Y. Jo, S. H. von Reuss, M. C. Genoff, P. W. Sternberg, and F. C. Schroeder. Targeted Metabolomics Reveals a Male Pheromone and Sex-Specific Ascaroside Biosynthesis in *Caenorhabditis elegans*. *ACS Chemical Biology*, 7(8):1321–1325, Aug. 2012.
- [277] N. Jaitly, A. Mayampurath, K. Littlefield, J. N. Adkins, G. A. Anderson, and R. D. Smith. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics*, 10(1):87, 2009.
- [278] A. T. James and A. J. P. Martin. Gas-liquid partition chromatography; the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid. *The Biochemical journal*, 50(5):679–690, Mar. 1952.
- [279] N. Jamshidi and B. O. Palsson. Formulating genome-scale kinetic models in the post-genome era. *Molecular Systems Biology*, 4:171, 2008.
- [280] G. Jansen, K. L. Thijssen, P. Werner, M. van der Horst, E. Hazendonk, and R. H. Plasterk. The complete family of genes encoding G proteins of *Caenorhabditis elegans*. *Nature Genetics*, 21(4):414–419, Apr. 1999.

- [281] P. Jatlow. Correction of Urine Cotinine Concentrations for Creatinine Excretion: Is It Useful? *Clinical Chemistry*, 49(11):1932–1934, Nov. 2003.
- [282] K. Jia, D. Chen, and D. L. Riddle. The TOR pathway interacts with the insulin signaling pathway to regulate *C. elegans* larval development, metabolism and life span. *Development*, 131(16):3897–3906, Aug. 2004.
- [283] K. Jia and B. Levine. Autophagy is required for dietary restriction-mediated life span extension in *C. elegans*. *Autophagy*, 3(6):597–599, Nov. 2007.
- [284] E. G. Johnson and A. O. Nier. Phys. Rev. 91, 10 (1953): Angular Aberrations in Sector Shaped Electromagnetic Lenses for Focusing Beams of Charged Particles. *Physical Review*, 1953.
- [285] H. E. Johnson, A. J. Lloyd, L. A. Mur, A. R. Smith, and D. R. Causton. The application of MANOVA to analyse Arabidopsis thaliana metabolomic data from factorially designed experiments. *Metabolomics*, 3(4):517–530, 2007.
- [286] S. C. Johnson, P. S. Rabinovitch, and M. Kaeberlein. mTOR is a key modulator of ageing and age-related disease. *Nature*, 493(7432):338–345, Jan. 2013.
- [287] T. Johnson, D. Wu, P. Tedesco, S. Dames, and J. Vaupel. Age-specific demographic profiles of longevity mutants in *Caenorhabditis elegans* show segmental effects. *The Journal of Gerontology*, 56A(8):331–339, 2001.
- [288] O. A. H. Jones, S. C. Swain, C. Svendsen, J. L. Griffin, S. R. Sturzenbaum, and D. J. Spurgeon. Potential New Method of Mixture Effects Testing Using Metabolomics and *Caenorhabditis elegans*. *Journal of Proteome Research*, 11(2):1446–1453, Feb. 2012.
- [289] S. Jozefczuk, S. Klie, G. Catchpole, J. Szymanski, A. Cuadros-Inostroza, D. Steinhäuser, J. Selbig, and L. Willmitzer. Metabolomic and transcriptomic stress response of Escherichia coli. *Molecular Systems Biology*, 364(6), 2010.
- [290] M. Kaeberlein. Lessons on longevity from budding yeast. *Nature*, 464(7288):513–519, Mar. 2010.
- [291] M. Kaeberlein, M. McVey, and L. Guarente. The SIR2/3/4 complex and SIR2 alone promote longevity in Saccharomyces cerevisiae by two different mechanisms. *Genes & Development*, 13(19):2570–2580, Oct. 1999.
- [292] T. L. Kaeberlein, E. D. Smith, M. Tsuchiya, K. L. Welton, J. H. Thomas, S. Fields, B. K. Kennedy, and M. Kaeberlein. Lifespan extension in *Caenorhabditis elegans* by complete removal of food. *Aging Cell*, 5(6):487–494, Dec. 2006.
- [293] R. Kamath and J. Ahringer. Genome-wide RNAi screening in *Caenorhabditis elegans*. *Methods*, 30(4):313–321, 2003.

- [294] R. Kamath, A. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, D. Welchman, P. Zipperlen, and J. Ahringer. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, 421(6920):231–237, 2003.
- [295] A. Kamleh, M. P. Barrett, D. Wildridge, R. J. S. Burchmore, R. A. Scheltema, and D. G. Watson. Metabolomic profiling using Orbitrap Fourier transform mass spectrometry with hydrophilic interaction chromatography: a method with wide applicability to analysis of biomolecules. *Rapid Communications In Mass Spectrometry*, 22(12):1912–1918, 2008.
- [296] M. A. Kamleh, T. M. D. Ebbels, K. Spagou, P. Masson, and E. J. Want. Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. *Analytical Chemistry*, 84(6):2670–2677, Mar. 2012.
- [297] Y. Kanfi, S. Naiman, G. Amir, V. Peshti, G. Zinman, L. Nahum, Z. Bar-Joseph, and H. Y. Cohen. The sirtuin SIRT6 regulates lifespan in male mice. *Nature*, 483(7388):218–221, Feb. 2012.
- [298] P. Kapahi, D. Chen, A. N. Rogers, S. D. Katewa, P. W.-L. Li, E. L. Thomas, and L. Kockel. With TOR, Less Is More: A Key Role for the Conserved Nutrient-Sensing TOR Pathway in Aging. *Cell Metabolism*, 11(6):453–465, June 2010.
- [299] F. Kaplan, D. V. Badri, C. Zachariah, R. Ajredini, F. J. Sandoval, S. Roje, L. H. Levine, F. Zhang, S. L. Robinette, H. T. Alborn, W. Zhao, M. Stadler, R. Nimalendran, A. T. Dossey, R. Brüscheweiler, J. M. Vivanco, and A. S. Edison. Bacterial attraction and quorum sensing inhibition in *Caenorhabditis elegans* exudates. *Journal of Chemical Ecology*, 35(8):878–892, Aug. 2009.
- [300] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons - Analytical Chemistry (ACS Publications). *Analytical Chemistry*, 60(20):2299–2301, Oct. 1988.
- [301] J. Kast, M. Gentzel, M. Wilm, and K. Richardson. Noise filtering techniques for electrospray quadrupole time of flight mass spectra. *Journal Of The American Society For Mass Spectrometry*, 14(7):766–776, July 2003.
- [302] M. Katajamaa, J. Miettinen, and M. Oresic. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics (Oxford, England)*, 22(5):634–636, Mar. 2006.
- [303] E.-B. Kayser, M. M. Sedensky, and P. G. Morgan. The effects of complex I function and oxidative damage on lifespan and anesthetic sensitivity in *Caenorhabditis elegans*. *Mechanisms Of Ageing And Development*, 125(6):455–464, June 2004.

- [304] P. Kebarle and U. H. Verkerk. Electrospray: From ions in solution to ions in the gas phase, what we know now. *Proteomics*, 28(6):898–917, Nov. 2009.
- [305] P. A. Keifer. Chemical-shift referencing and resolution stability in gradient LC-NMR (acetonitrile:water). *Journal Of Magnetic Resonance*, 199(1):75–87, July 2009.
- [306] B. O. Keller, J. Sui, A. B. Young, and R. M. Whittall. Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta*, 627(1):71–81, Oct. 2008.
- [307] B. Kennedy. The genetics of ageing: insight from genome-wide approaches in invertebrate model organisms. *Journal of Internal Medicine*, 263(2):142–152, 2008.
- [308] B. K. Kennedy and M. Kaeblerlein. Hot topics in aging research: protein translation, 2009. *Aging Cell*, 8(6):617–623, 2009.
- [309] C. Kenyon. The plasticity of aging: Insights from long-lived mutants. *Cell*, 120(4):449–460, 2005.
- [310] C. Kenyon, J. Chang, E. Gensch, A. Rudner, and R. Tabtiang. A *C. elegans* mutant that lives twice as long as wild type. *Nature*, 366(6454):461–464, Dec. 1993.
- [311] C. J. Kenyon. The genetics of ageing. *Nature*, 464(7288):504–512, 2010.
- [312] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, 39(Database issue):D583–90, Jan. 2011.
- [313] I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martinez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñiz-Rascado, Q. Ong, S. Paley, I. Schröder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen, and P. D. Karp. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Research*, 41(Database issue):D605–12, Jan. 2013.
- [314] H. Keun, T. Ebbels, H. Antti, M. Bollard, O. Beckonert, G. Schlotterbeck, H. Senn, U. Niederhauser, E. Holmes, J. Lindon, and J. Nicholson. Analytical reproducibility in 1H NMR-based metabonomic urinalysis. *Chemical Research In Toxicology*, 2002.
- [315] P. Kiefer, U. Schmitt, and J. A. Vorholt. eMZed: an open source framework in Python for rapid and interactive development of LC/MS data analysis workflows. *Bioinformatics (Oxford, England)*, 29(7):963–964, Apr. 2013.

- [316] J. Kikuchi, K. Shinozaki, and T. Hirayama. Stable isotope labeling of *Arabidopsis thaliana* for an NMR-based metabolomics approach. *Plant & Cell Physiology*, 45(8):1099–1104, Aug. 2004.
- [317] K. D. Kimura, H. A. Tissenbaum, Y. Liu, and G. Ruvkun. daf-2, an insulin receptor-like gene that regulates longevity and diapause in *Caenorhabditis elegans*. *Science (New York, NY)*, 277(5328):942–946, Aug. 1997.
- [318] Y. Kimura, Y. Sugiura, T. Hayasaka, and M. Setou. Visualizing the spatial distribution of biomolecules in *C. elegans* by imaging mass Spectrometry. *The Worm Breeder's Gazette*, 18(1):4, 2009.
- [319] T. Kind and O. Fiehn. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7:234, 2006.
- [320] T. Kind and O. Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8:105, 2007.
- [321] T. Kind and O. Fiehn. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews*, 2(1-4):23–60, Dec. 2010.
- [322] T. Kind, K.-H. Liu, D. Y. Lee, B. Defelice, J. K. Meissen, and O. Fiehn. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nature Methods*, 10(8):755–758, June 2013.
- [323] T. Kind, G. Wohlgemuth, D. Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, and O. Fiehn. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry*, 81(24):10038–10048, Dec. 2009.
- [324] D. N. Kirk, H. C. Toms, C. Douglas, K. A. White, K. E. Smith, S. Latif, and R. W. P. Hubbard. A survey of the high-field ^1H NMR spectra of the steroid hormones, their hydroxylated derivatives, and related compounds. *J. Chem. Soc., Perkin Trans. 2*, 0(9):1567–1594, 1990.
- [325] T. B. Kirkwood. Evolution of ageing. *Nature*, 270(5635):301–304, Nov. 1977.
- [326] T. B. L. Kirkwood. Understanding the odd science of aging. *Cell*, 120(4):437–447, Feb. 2005.
- [327] T. B. L. Kirkwood, R. J. Boys, C. S. Gillespie, C. J. Proctor, D. P. Shanley, and D. J. Wilkinson. Towards an e-biology of ageing: integrating theory and data. *Nature Reviews Molecular Cell Biology*, 4(3):243–249, Mar. 2003.
- [328] H. Kitano. Systems biology: a brief overview. *Science (New York, NY)*, 295(5560):1662–1664, Mar. 2002.

- [329] R. Kleemann, M. Van Erk, L. Verschuren, A. Van Den Hoek, M. Koek, P. Wielinga, A. Jie, L. Pellis, I. Bobeldijk-Pastorova, and T. Kelder. Time-resolved and tissue-specific systems analysis of the pathogenesis of insulin resistance. *PloS one*, 5(1):840–846, 2010.
- [330] D. J. Klionsky. Autophagy: from phenomenology to molecular understanding in less than a decade. *Nature Reviews Molecular Cell Biology*, 8(11):931–937, Nov. 2007.
- [331] C. Knight, M. Patel, R. Azevedo, and A. Leroi. A novel mode of ecdysozoan growth in *Caenorhabditis elegans*. *Evolution & Development*, 4(1):16–27, 2002.
- [332] Y. Koh, K. K. Pasikanti, C. W. Yap, and E. C. Y. Chan. Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data. *Journal Of Chromatography A*, 1217(52):8308–8316, Dec. 2010.
- [333] P. Koivunen, S. Lee, C. G. Duncan, G. Lopez, G. Lu, S. Ramkissoon, J. A. Losman, P. Joensuu, U. Bergmann, S. Gross, J. Travins, S. Weiss, R. Looper, K. L. Ligon, R. G. W. Verhaak, H. Yan, and W. G. Kaelin. Transformation by the (R)-enantiomer of 2-hydroxyglutarate linked to EGLN activation. *Nature*, 483(7390):484–488, Mar. 2012.
- [334] L. Konermann, E. Ahadi, A. D. Rodriguez, and S. Vahidi. Unraveling the mechanism of electrospray ionization. *Analytical Chemistry*, 85(1):2–9, Jan. 2013.
- [335] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmüller, P. Dörmann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, and D. Steinhauser. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics (Oxford, England)*, 21(8):1635–1638, Apr. 2005.
- [336] E. Kováts. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helvetica Chimica Acta*, 41(7):1915–1932, 1958.
- [337] J. Krijgsveld, R. F. Ketting, T. Mahmoudi, J. Johansen, M. Artal-Sanz, C. P. Verrijzer, R. H. A. Plasterk, and A. J. R. Heck. Metabolic labeling of *C. elegans* and *D. melanogaster* for quantitative proteomics. *Nature Biotechnology*, 21(8):927–931, Aug. 2003.
- [338] C. Kuhl, R. Tautenhahn, C. Bottcher, T. R. Larson, and S. Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. *Analytical Chemistry*, 84(1):283–289, Nov. 2011.
- [339] T. V. Kurzchalia and S. Ward. Why do worms need cholesterol? *Nature Cell Biology*, 5(8):684–688, Aug. 2003.

- [340] E.-S. Kwon, S. D. Narasimhan, K. Yen, and H. A. Tissenbaum. A new DAF-16 isoform regulates longevity. *Nature*, 466(7305):498–502, July 2010.
- [341] B. Lakowski and S. Hekimi. Determination of Life-Span in *Caenorhabditis elegans* by Four Clock Genes. *Science (New York, NY)*, 272(5264):1010–1013, May 1996.
- [342] B. Lakowski and S. Hekimi. The genetics of caloric restriction in *Caenorhabditis elegans*. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 95(22):13091–13096, 1998.
- [343] D. Lange, C. Storment, C. Conley, and G. Kovacs. A microfluidic shadow imaging system for the study of the nematode *Caenorhabditis elegans* in space. *Sensors And Actuators B-Chemical*, 107(2):904–914, 2005.
- [344] I. Langmuir. The Arrangement of Electrons and Atoms in Molecules . *Journal of the American Chemical Society*, 41(6):868–934, June 1919.
- [345] L. R. Lapierre, S. Gelino, A. Meléndez, and M. Hansen. Autophagy and Lipid Metabolism Coordinately Modulate Life Span in Germline-less *C. elegans*. *Current Biology*, 21(18):1507–1514, Sept. 2011.
- [346] L. R. Lapierre and M. Hansen. Lessons from *C. elegans*: signaling pathways for longevity. *Trends in Endocrinology & Metabolism*, 23(12):637–644, Dec. 2012.
- [347] M. Laplante and D. M. Sabatini. An Emerging Role of mTOR in Lipid Biosynthesis. *Current Biology*, 19(22):R1046–R1052, Dec. 2009.
- [348] M. Laplante and D. M. Sabatini. Regulation of mTORC1 and its impact on gene expression at a glance. *Journal Of Cell Science*, 126(Pt 8):1713–1719, Apr. 2013.
- [349] J. Le Belle, N. Harris, S. Williams, and K. Bhakoo. A comparison of cell and tissue extraction techniques using high-resolution H-1-NMR spectroscopy. *Nmr In Biomedicine*, 15(1):37–44, 2002.
- [350] E.-Y. Lee, Y.-H. Shim, D. J. Chitwood, S. B. Hwang, J. Lee, and Y.-K. Paik. Cholesterol-producing transgenic *Caenorhabditis elegans* lives longer due to newly acquired enhanced stress resistance. *Biochemical and Biophysical Research Communications*, 328(4):929–936, Mar. 2005.
- [351] S. Lee, R. Lee, A. Fraser, R. Kamath, J. Ahringer, and G. Ruvkun. A systematic RNAi screen identifies a critical role for mitochondria in C-elegans longevity. *Nature Genetics*, 33(1):40–48, 2003.
- [352] S.-J. Lee, A. B. Hwang, and C. Kenyon. Inhibition of respiration extends *C. elegans* life span via reactive oxygen species that increase HIF-1 activity. *Current Biology*, 20(23):2131–2136, Dec. 2010.

- [353] S.-J. Lee, C. T. Murphy, and C. Kenyon. Glucose shortens the life span of *C. elegans* by downregulating DAF-16/FOXO activity and aquaporin gene expression. *Cell Metabolism*, 10(5):379–391, Nov. 2009.
- [354] S. S. Lee, S. Kennedy, A. C. Tolonen, and G. Ruvkun. DAF-16 target genes that control *C. elegans* life-span and metabolism. *Science (New York, NY)*, 300(5619):644–647, Apr. 2003.
- [355] S. S. Lee and F. C. Schroeder. Steroids as Central Regulators of Organismal Development and Lifespan. *PLoS Biology*, 10(4):e1001307, Apr. 2012.
- [356] S. F. Leiser, A. Begun, and M. Kaeberlein. HIF-1 modulates longevity and healthspan in a temperature-dependent manner. *Aging Cell*, 10(2):318–326, Apr. 2011.
- [357] S. F. Leiser, M. Fletcher, A. Begun, and M. Kaeberlein. Life-Span Extension From Hypoxia in *Caenorhabditis elegans* Requires Both HIF-1 and DAF-16 and Is Antagonized by SKN-1. *The Journals of Gerontology Series A, Biological Sciences and Medical Sciences*, Feb. 2013.
- [358] A. Leroi. Molecular signals versus the Loi de Balancement. *Trends in Ecology & Evolution*, 16(1):24–29, Jan. 2001.
- [359] E. Levine and T. Hwa. Stochastic fluctuations in metabolic pathways. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 104(22):9224–9229, May 2007.
- [360] M. H. Levitt. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley, 2008.
- [361] I. A. Lewis, S. C. Schommer, and J. L. Markley. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magnetic Resonance In Chemistry*, 47 Suppl 1:S123–6, Dec. 2009.
- [362] W. Li, B. Gao, S.-M. Lee, K. Bennett, and D. Fang. RLE-1, an E3 ubiquitin ligase, regulates *C. elegans* aging by catalyzing DAF-16 polyubiquitination. *Developmental Cell*, 12(2):235–246, Feb. 2007.
- [363] C.-Y. Liao and B. K. Kennedy. Will the real aging Sirtuin please stand up? *Cell Research*, 22(8):1215–1217, Aug. 2012.
- [364] M. Liebeke and J. G. Bundy. Tissue disruption and extraction methods for metabolic profiling of an invertebrate sentinel species. *Metabolomics*, Oct. 2011.
- [365] M. Liebeke, J. Hao, T. M. D. Ebbels, and J. G. Bundy. Combining spectral ordering with peak fitting for one-dimensional NMR quantitative metabolomics. *Analytical Chemistry*, 85(9):4605–4612, May 2013.

- [366] M. Liebeke, H. Meyer, S. Donat, K. Ohlsen, and M. Lalk. A Metabolomic View of *Staphylococcus aureus* and Its Ser/Thr Kinase and Phosphatase Deletion Mutants: Involvement in Cell Wall Biosynthesis. *Chemistry & Biology*, 17(8):820–830, Aug. 2010.
- [367] C. Y. Lin, H. Wu, R. S. Tjeerdema, and M. R. Viant. Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. *Metabolomics*, 3(1):55–67, 2007.
- [368] K. Lin, J. B. Dorman, A. Rodan, and C. Kenyon. daf-16: An HNF-3/forkhead family member that can function to double the life-span of *Caenorhabditis elegans*. *Science (New York, NY)*, 278(5341):1319–1322, Nov. 1997.
- [369] S.-J. Lin, M. Kaeberlein, A. A. Andalis, L. A. Sturtz, P.-A. Defossez, V. C. Culotta, G. R. Fink, and L. Guarente. Calorie restriction extends *Saccharomyces cerevisiae* lifespan by increasing respiration. *Nature*, 418(6895):344–348, July 2002.
- [370] N. J. Linford, T.-H. Kuo, T. P. Chan, and S. D. Pletcher. Sensory Perception and Aging in Model Systems: From the Outside In. *Annual Review of Cell and Developmental Biology*, 27:759–785, 2011.
- [371] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer. The Hallmarks of Aging. *Cell*, 153(6):1194–1217, June 2013.
- [372] J.-A. Losman, R. E. Looper, P. Koivunen, S. Lee, R. K. Schneider, C. McMahon, G. S. Cowley, D. E. Root, B. L. Ebert, and W. G. Kaelin. (R)-2-hydroxyglutarate is sufficient to promote leukemogenesis and its effects are reversible. *Science (New York, NY)*, 339(6127):1621–1625, Mar. 2013.
- [373] W. J. Lough and I. W. Wainer. *High-performance Liquid Chromatography: Fundamental Principles and Practice*. Blackie Academic & Professional, 1995.
- [374] C. Lu, P. S. Ward, G. S. Kapoor, D. Rohle, S. Turcan, O. Abdel-Wahab, C. R. Edwards, R. Khanin, M. E. Figueroa, A. Melnick, K. E. Wellen, D. M. O’Rourke, S. L. Berger, T. A. Chan, R. L. Levine, I. K. Mellinghoff, and C. B. Thompson. IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature*, 483(7390):474–478, Mar. 2012.
- [375] W. Lu, M. F. Clasquin, E. Melamud, D. Amador-Noguez, A. A. Caudy, and J. D. Rabinowitz. Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Analytical Chemistry*, 82(8):3212–3221, Apr. 2010.
- [376] C. Ludwig and M. R. Viant. Two-dimensional J-resolved NMR Spectroscopy: Review of a Key Methodology in the Metabolomics Toolbox. *Phytochemical Analysis*, 21(1):22–32, 2010.

- [377] J. Lund, P. Tedesco, K. Duke, J. Wang, and S. Kim. Transcriptional profile of aging in *C. elegans*. *Current Biology*, 12:1566–1573, 2002.
- [378] B. Luo, K. Groenke, R. Takors, C. Wandrey, and M. Oldiges. Simultaneous determination of multiple intracellular metabolites in glycolysis, pentose phosphate pathway and tricarboxylic acid cycle by liquid chromatography-mass spectrometry. *Journal Of Chromatography A*, 1147(2):153–164, Apr. 2007.
- [379] N. W. Lutz and P. J. Cozzzone. Multiparametric optimization of (31)P NMR spectroscopic analysis of phospholipids in crude tissue extracts. 2. Line width and spectral resolution. *Analytical Chemistry*, 82(13):5441–5446, July 2010.
- [380] N. W. Lutz and P. J. Cozzzone. Multiparametric Optimization of P-31 NMR Spectroscopic Analysis of Phospholipids in Crude Tissue Extracts. 1. Chemical Shift and Signal Separation. *Analytical Chemistry*, 82(13):5433–5440, 2010.
- [381] H. Ma, L. Jiang, W. Shi, J. Qin, and B. Lin. A programmable microvalve-based microfluidic array for characterization of neurotoxin-induced responses of individual *C. elegans*. *Biomicrofluidics*, 3(4):044114, 2009.
- [382] L. T. MacNeil, E. Watson, H. E. Arda, L. J. Zhu, and A. J. M. Walhout. Diet-Induced Developmental Acceleration Independent of TOR and Insulin in *C. elegans*. *Cell*, 153(1):240–252, Mar. 2013.
- [383] F. Madeo, N. Tavernarakis, and G. Kroemer. Can autophagy promote longevity? *Nature Cell Biology*, 12(9):842–846, Sept. 2010.
- [384] R. Madsen, T. Lundstedt, and J. Trygg. Chemometrics in metabolomics—A review in human disease diagnosis. *International Journal of Pharmaceutics*, 659(1-2):23–33, Feb. 2010.
- [385] W. Mair, S. H. Panowski, R. J. Shaw, and A. Dillin. Optimizing dietary restriction for genetic epistasis analysis and gene discovery in *C. elegans*. *PloS one*, 4(2):e4535, 2009.
- [386] H. Y. Mak, L. S. Nelson, M. Basson, C. D. Johnson, and G. Ruvkun. Polygenic control of *Caenorhabditis elegans* fat storage. *Nature Genetics*, 38(3):363–368, Feb. 2006.
- [387] A. Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry*, 72(6):1156–1162, Mar. 2000.
- [388] C. R. Mallet, Z. Lu, and J. R. Mazzeo. A study of ion suppression effects in electrospray ionization from mobile phase additives and solid-phase extracts. *Rapid Communications in Mass Spectrometry*, 18(1):49–58, 2004.

- [389] J. Malm, D. Giannaras, M. O. Riehle, N. Gadegaard, and P. Sjövall. Fixation and Drying Protocols for the Preparation of Cell Samples for Time-of-Flight Secondary Ion Mass Spectrometry Analysis. *Analytical Chemistry*, 81(17):7197–7205, 2009.
- [390] B. A. Mamyrin, V. I. Karataev, and D. V. Shmikk. The mass-reflectron, a new non-magnetic time-of-flight mass spectrometer with high resolution. *Soviet Phys JETP*, 37:45, 1973.
- [391] P. K. Mandal and A. Majumdar. A comprehensive discussion of HSQC and HMQC pulse sequences. *Concepts in Magnetic Resonance*, 20A(1):1–23, 2004.
- [392] F.-P. J. Martin, B. Spanier, S. Collino, I. Montoliu, C. Kolmeder, P. Giesbertz, M. Affolter, M. Kussmann, H. Daniel, S. Kochhar, and S. Rezzi. Metabotyping of *Caenorhabditis elegans* and their culture media revealed unique metabolic phenotypes associated to amino acid deficiency and insulin-like signaling. *Journal of Proteome Research*, 10(3):990–1003, Jan. 2012.
- [393] S. Mas, R. Perez, R. Martinez-Pinna, J. Egido, and F. Vivanco. Cluster TOF-SIMS imaging: a new light for in situ metabolomics? *Proteomics*, 8(18):3735–3745, Sept. 2008.
- [394] P. Masson, A. C. Alves, T. M. D. Ebbels, J. K. Nicholson, and E. J. Want. Optimization and evaluation of metabolite extraction protocols for untargeted metabolic profiling of liver samples by UPLC-MS. *Analytical Chemistry*, 82(18):7779–7786, Sept. 2010.
- [395] J. Mattauch and R. Herzog. Über einen neuen Massenspektrographen. *Zeitschrift für Physik A Hadrons and Nuclei*, 1934.
- [396] V. Matyash, G. Liebisch, T. V. Kurzchalia, A. Shevchenko, and D. Schwudke. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal Of Lipid Research*, 49(5):1137–1146, May 2008.
- [397] C. M. McCay and M. F. Crowell. Prolonging the Life Span. *The Scientific Monthly*, 39:405–414, Nov. 1934.
- [398] M. A. McCormick, S.-y. Tsai, and B. K. Kennedy. TOR and ageing: a complex pathway for a complex process. *Philosophical Transactions Of The Royal Society Of London Series B-Biological Sciences*, 366(1561):17–27, Jan. 2011.
- [399] J. McDonald and G. Whitesides. Poly (dimethylsiloxane) as a material for fabricating microfluidic devices. *Acc. Chem. Res*, 35(7):491–499, 2002.
- [400] J. McElwee, K. Bubb, and J. H. Thomas. Transcriptional outputs of the *Caenorhabditis elegans* forkhead protein DAF-16. *Aging Cell*, 2(2):111–121, Apr. 2003.

- [401] M. McGue, J. Vaupel, N. Holm, and B. Harvald. Longevity is moderately heritable in a sample of Danish twins born 1870-1880. *The Journal of Biological Sciences*, 48(6):237-244, 1993.
- [402] F. W. McLafferty. Mass spectrometry in chemical research and production. *Applied Spectroscopy*, 11(4):148-156, 1957.
- [403] F. W. McLafferty. Mass Spectrometric Analysis. Molecular Rearrangements. *Analytical Chemistry*, 31(1):82-87, Jan. 1959.
- [404] F. W. McLafferty, M. Y. Zhang, D. B. Stauffer, and S. Y. Loh. Comparison of algorithms and databases for matching unknown mass spectra. *Journal Of The American Society For Mass Spectrometry*, 9(1):92-95, 1998.
- [405] A. D. McNaught and A. Wilkinson. *IUPAC Compendium of Chemical Terminology*. WileyBlackwell, 2rev ed edition, 1997.
- [406] C. M. McQuaw, L. Zheng, A. G. Ewing, and N. Winograd. Localization of sphingomyelin in cholesterol domains by imaging mass spectrometry. *Langmuir*, 23(10):5645-5650, May 2007.
- [407] P. Medwar. *An Unsolved Problem of Biology*. HK Lewis and Co., 1952.
- [408] R. Mehta, K. A. Steinkraus, G. L. Sutphin, F. J. Ramos, L. S. Shamieh, A. Huh, C. Davis, D. Chandler-Brown, and M. Kaeberlein. Proteasomal regulation of the hypoxic response modulates aging in *C. elegans*. *Science (New York, NY)*, 324(5931):1196-1198, May 2009.
- [409] E. Melamud, L. Vastag, and J. D. Rabinowitz. Metabolomic Analysis and Visualization Engine for LC-MS Data. *Analytical Chemistry*, 82(23):9818-9826, Dec. 2010.
- [410] A. Meléndez, Z. Tallóczy, M. Seaman, E.-L. Eskelinen, D. H. Hall, and B. Levine. Autophagy genes are essential for dauer development and life-span extension in *C. elegans*. *Science (New York, NY)*, 301(5638):1387-1391, Sept. 2003.
- [411] E. M. Mercken, S. D. Crosby, D. W. Lammington, L. J. Jeabailey, S. Krzysik-Walker, D. Villareal, M. Capri, C. Franceschi, Y. Zhang, K. Becker, D. M. Sabatini, R. de Cabo, and L. Fontana. Calorie restriction in humans inhibits the PI3K/AKT pathway and induces a younger transcription profile. *Aging Cell*, Apr. 2013.
- [412] A. Merkel. *Untersuchung des Mechanismus von Zerfallsreaktionen mit einfachem Bindungsbruch und Berechnung ihrer Geschwindigkeitskonstanten auf der Grundlage quantenchemischer und statistischer Methoden*. PhD thesis, Central Institute for Physical Chemistry, 1986.

- [413] M. Merris, J. Kraeft, G. S. Tint, and J. Lenard. Long-term effects of sterol depletion in *C. elegans*: sterol content of synchronized wild-type and mutant populations. *The Journal of Lipid Research*, 45(11):2044–2051, Nov. 2004.
- [414] H. Meyer, M. Liebeke, and M. Lalk. A protocol for the investigation of the intracellular *Staphylococcus aureus* metabolome. *Analytical Biochemistry*, 401(2):250–259, June 2010.
- [415] N. Minois, P. Sykacek, B. Godsey, and D. Kreil. RNA Interference in Ageing Research—A Mini-Review. *Gerontology*, 56:496–506, 2010.
- [416] A. A. Miranda, Y.-A. Borgne, and G. Bontempi. New Routes from Minimal Approximation Error to Principal Components. *Neural Processing Letters*, 27(3):197–207, Jan. 2008.
- [417] R. J. Mishur and S. L. Rea. Applications of mass spectrometry to metabolomics and metabonomics: detection of biomarkers of aging and of age-related diseases. *Mass Spectrometry Reviews*, 31(1):70–95, Jan. 2012.
- [418] M. Miyamoto and S. Sasakawa. Effects of autoclave sterilization on the physical properties of storage bags and granulocyte function. *Vox Sanguinis*, 54(2):74–77, 1988.
- [419] I. Molnar. Computerized design of separation strategies by reversed-phase liquid chromatography: development of DryLab software. *Journal Of Chromatography A*, 965(1-2):175–194, Aug. 2002.
- [420] H. R. Morris, M. Panico, M. Barber, R. S. Bordoli, R. D. Sedgwick, and A. Tyler. Fast atom bombardment: a new mass spectrometric method for peptide sequence analysis. *Biochemical and Biophysical Research Communications*, 101(2):623–631, July 1981.
- [421] H. R. Morris, T. Paxton, A. Dell, J. Langhorne, M. Berg, R. S. Bordoli, J. Hoyes, and R. H. Bateman. High sensitivity collisionally- activated decomposition tandem mass spectrometry on a novel quadrupole orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Communications In Mass Spectrometry*, 10(8):889–896, 1996.
- [422] H. R. Morris, T. Paxton, M. Panico, R. McDowell, and A. Dell. A novel geometry mass spectrometer, the Q-TOF, for low-femtomole/attomole-range biopolymer sequencing. *Journal of Protein Chemistry*, 16(5):469–479, 1997.
- [423] J. Z. Morris, H. A. Tissenbaum, and G. Ruvkun. A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in *Caenorhabditis elegans*. *Nature*, 382(6591):536–539, Aug. 1996.

- [424] E. Morselli, L. Galluzzi, O. Kepp, A. Criollo, M. Maiuri, N. Tavernarakis, F. Madeo, and G. Kroemer. Autophagy mediates pharmacological lifespan extension by spermidine and resveratrol. *Aging*, 1(12):961–970, 2009.
- [425] E. Morselli, M. C. Maiuri, M. Markaki, E. Megalou, A. Pasparaki, K. Palikaras, A. Criollo, L. Galluzzi, S. A. Malik, I. Vitale, M. Michaud, F. Madeo, N. Tavernarakis, and G. Kroemer. Caloric restriction and resveratrol promote longevity through the Sirtuin-1-dependent induction of autophagy. *Cell Death & Disease*, 1:e10, 2010.
- [426] E. Morselli, M. C. Maiuri, M. Markaki, E. Megalou, A. Pasparaki, K. Palikaras, A. Criollo, L. Galluzzi, S. A. Malik, I. Vitale, M. Michaud, F. Madeo, N. Tavernarakis, and G. Kroemer. The life span-prolonging effect of sirtuin-1 is mediated by autophagy. *Autophagy*, 6(1):186–188, Jan. 2010.
- [427] D. L. Motola, C. L. Cummins, V. Rottiers, K. K. Sharma, T. Li, Y. Li, K. Suino-Powell, H. E. Xu, R. J. Auchus, and A. Antebi. Identification of Ligands for DAF-12 that Govern Dauer Formation and Reproduction in *C. elegans*. *Cell*, 124(6):1209–1223, 2006.
- [428] MSI Board Members, S.-A. Sansone, T. Fan, R. Goodacre, J. L. Griffin, N. W. Hardy, R. Kaddurah-Daouk, B. S. Kristal, J. Lindon, P. Mendes, N. Morrison, B. Nikolau, D. Robertson, L. W. Sumner, C. Taylor, M. van der Werf, B. van Ommen, and O. Fiehn. The metabolomics standards initiative. *Nature Biotechnology*, 25(8):846–848, Aug. 2007.
- [429] R.-U. Müller, F. Fabretti, S. Zank, V. Burst, T. Benzing, and B. Schermer. The von Hippel Lindau tumor suppressor limits longevity. *Journal of the American Society of Nephrology*, 20(12):2513–2517, Dec. 2009.
- [430] M. S. B. Munson and F. H. Field. . *Analytical Chemistry*, 88(12):2621–2630, June 1966.
- [431] S. Murakami and T. E. Johnson. A genetic pathway conferring life extension and resistance to UV stress in *Caenorhabditis elegans*. *Genetics*, 143(3):1207–1218, July 1996.
- [432] C. T. Murphy, S. A. McCarroll, C. I. Bargmann, A. Fraser, R. S. Kamath, J. Ahringer, H. Li, and C. Kenyon. Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature*, 424(6946):277–283, July 2003.
- [433] P. Murray, S. A. L. Hayward, G. G. Govan, A. Y. Gracey, and A. R. Cossins. An explicit test of the phospholipid saturation hypothesis of acquired cold tolerance in *Caenorhabditis elegans*. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 104(13):5489–5494, Mar. 2007.

- [434] R. Mylonas, Y. Mauron, A. Masselot, P.-A. Binz, N. Budin, M. Fathi, V. Viette, D. F. Hochstrasser, and F. Lisacek. X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry. *Analytical Chemistry*, 81(18):7604–7610, Sept. 2009.
- [435] R. Nadon and J. Shoemaker. Statistical issues with microarrays: processing and analysis. *Trends in Genetics*, 18(5):265–271, May 2002.
- [436] J. H. Naismith, M.-F. Giraud, G. A. Leonard, R. A. Field, and C. Berlind. RmlC, the third enzyme of dTDP-L-rhamnose pathway, is a new class of epimerase - Nature Structural & Molecular Biology. *Nature Structural Biology*, 7(5):398–402, May 2000.
- [437] S. D. Narasimhan, K. Yen, and H. A. Tissenbaum. Converging Pathways in Lifespan Regulation. *Current Biology*, 19(15):R657–R666, Aug. 2009.
- [438] P. Narbonne and R. Roy. *Caenorhabditis elegans* dauers need LKB1/AMPK to ration lipid reserves and ensure long-term survival. *Nature*, 457(7226):210–214, Jan. 2009.
- [439] Z. Ni and S. Lee. RNAi screens to identify components of gene networks that modulate aging in *Caenorhabditis elegans*. *Briefings in Functional Genomics*, 9(1):53–64, 2010.
- [440] J. Nicholson, J. Connelly, J. Lindon, and E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*, 1(2):153–161, 2002.
- [441] J. Nicholson, E. Holmes, J. Lindon, and I. Wilson. The challenges of modeling mammalian biocomplexity. *Nature Biotechnology*, 22(10):1268–1274, 2004.
- [442] J. K. Nicholson, P. J. Foxall, M. Spraul, R. D. Farrant, and J. C. Lindon. 750 MHz ^1H and ^1H - ^{13}C NMR spectroscopy of human blood plasma. *Analytical Chemistry*, 67(5):793–811, Mar. 1995.
- [443] J. K. Nicholson and J. C. Lindon. Systems biology - Metabonomics. *Nature*, 455(7216):1054–1056, 2008.
- [444] J. K. Nicholson, J. C. Lindon, and E. Holmes. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29(11):1181–1189, Nov. 1999.
- [445] W. M. Niessen. Advances in instrumentation in liquid chromatography-mass spectrometry and related liquid-introduction techniques. *Journal Of Chromatography A*, 794(1-2):407–435, Jan. 1998.

- [446] A. Nordstrom, G. O'Maille, C. Qin, and G. Siuzdak. Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. *Analytical Chemistry*, 78(10):3289–3295, May 2006.
- [447] T. R. Northen, O. Yanes, M. T. Northen, D. Marrinucci, W. Uritboonthai, J. Apon, S. L. Golledge, A. Nordstrom, and G. Siuzdak. Clathrate nanostructures for mass spectrometry. *Nature*, 449(7165):1033–U3, 2007.
- [448] H. Nygren, B. Hagenhoff, P. Malmberg, M. Nilsson, and K. Richter. Bioimaging TOF-SIMS: High resolution 3D imaging of single cells. *Microscopy Research and Technique*, 70(11):969–974, Nov. 2007.
- [449] I. G. O'Brien, G. B. Cox, and F. Gibson. Biologically active compounds containing 2,3-duhydroxybenzoic acid and serine formed by *Escherichia coli*. *Biochimica et Biophysica Acta*, 201(3):453–460, Mar. 1970.
- [450] H. Ogata, S. Goto, K. Sato, and W. Fujibuchi. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.
- [451] R. J. Ogg, R. B. Kingsley, and J. S. Taylor. WET, a T1- and B1-Insensitive Water-Suppression Method for in Vivo Localized ¹H NMR Spectroscopy. *Journal of Magnetic Resonance, Series B*, 104(1):1–10, May 1994.
- [452] S. Ogg and G. Ruvkun. The *C. elegans* PTEN homolog, DAF-18, acts in the insulin receptor-like metabolic signaling pathway. *Molecular Cell*, 2(6):887–893, Dec. 1998.
- [453] S. W. Oh, A. Mukhopadhyay, B. L. Dixit, T. Raha, M. R. Green, and H. A. Tissenbaum. Identification of direct DAF-16 targets controlling longevity, metabolism and diapause by chromatin immunoprecipitation. *Nature Genetics*, 38(2):251–257, Feb. 2006.
- [454] S. W. Oh, A. Mukhopadhyay, N. Svrtikapa, F. Jiang, R. J. Davis, and H. A. Tissenbaum. JNK regulates lifespan in *Caenorhabditis elegans* by modulating nuclear translocation of forkhead transcription factor/DAF-16. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 102(12):4494–4499, Mar. 2005.
- [455] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz. Systematic functional analysis of the yeast genome. *Trends In Biotechnology*, 16(9):373–378, Sept. 1998.
- [456] A. Olsen, M. Vantipalli, and G. Lithgow. Lifespan extension of *Caenorhabditis elegans* following repeated mild hormetic heat treatments. *Biogerontology*, 7:221–230, 2006.
- [457] K. L. Olszewski, M. W. Mather, J. M. Morrissey, B. A. Garcia, A. B. Vaidya, J. D. Rabinowitz, and M. Llinas. Branched tricarboxylic acid metabolism in *Plasmodium falciparum*. *Nature*, 466(7307):774–U16, 2010.

- [458] E. J. O'Rourke and G. Ruvkun. MXL-3 and HLH-30 transcriptionally link lipolysis and autophagy to nutrient availability. *Nature Cell Biology*, 15(6):668–676, Apr. 2013.
- [459] S. G. Ostrowski, M. E. Kurczy, T. P. Roddy, N. Winograd, and A. G. Ewing. Secondary ion MS imaging to relatively quantify cholesterol in the membranes of individual cells from differentially treated populations. *Analytical Chemistry*, 79(10):3554–3560, May 2007.
- [460] J. Palter, A. Rogers, A. Olsen, and D. Chen. Inhibition of mRNA translation extends lifespan in *Caenorhabditis elegans*. *Aging*, 6(1):111–119, 2007.
- [461] S. H. Panowski, S. Wolff, H. Aguilaniu, J. Durieux, and A. Dillin. PHA-4/Foxa mediates diet-restriction-induced longevity of *C. elegans*. *Nature*, 447(7144):550–555, May 2007.
- [462] S. Paradis, M. Ailion, A. Toker, J. H. Thomas, and G. Ruvkun. A PDK1 homolog is necessary and sufficient to transduce AGE-1 PI3 kinase signals that regulate diapause in *Caenorhabditis elegans*. *Genes & Development*, 13(11):1438–1452, June 1999.
- [463] S. Paradis and G. Ruvkun. *Caenorhabditis elegans* Akt/PKB transduces insulin receptor-like signals from AGE-1 PI3 kinase to the DAF-16 transcription factor. *Genes & Development*, 12(16):2488–2498, Aug. 1998.
- [464] H. M. Parsons, C. Ludwig, U. L. Günther, and M. R. Viant. Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*, 8(1):234, 2007.
- [465] L. Partridge. Some highlights of research on aging with invertebrates, 2009. *Aging Cell*, 8(5):509–513, Sept. 2009.
- [466] G. J. Patti, H.-K. Woo, O. Yanes, L. Shriver, D. Thomas, W. Uritboonthai, J. V. Apon, R. Steenwyk, M. Manchester, and G. Siuzdak. Detection of carbohydrates and steroids by cation-enhanced nanostructure-initiator mass spectrometry (NIMS) for biofluid analysis and tissue imaging. *Analytical Chemistry*, 82(1):121–128, 2010.
- [467] W. Paul. Electromagnetic traps for charged and neutral particles (Nobel lecture). *Angewandte Chemie International Edition in English*, 29(7):739–748, 1990.
- [468] W. Paul and H. Steinwedel. Ein neues massenspektrometer ohne magnetfeld. *Zeitschrift Naturforschung Teil A*, 1953.
- [469] I. P. Pavlov and W. H. Thompson. *The work of the digestive glands*. Griffin's medical series. C. Griffin, 1902.
- [470] R. Pearl. *The rate of living: being an account of some experimental studies on the biology of life duration*. A.A. Knopf, 1928.

- [471] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- [472] P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, 22(11):1459–1466, Nov. 2004.
- [473] H. E. Pence and A. Williams. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education*, 87(11):1123–1124, Nov. 2010.
- [474] C. L. Perez and M. R. Van Gilst. A ^{13}C isotope labeling strategy reveals the influence of insulin signaling on lipogenesis in *C. elegans*. *Cell Metabolism*, 8(3):266–274, Sept. 2008.
- [475] T. T. Perls, J. Wilmoth, R. Levenson, M. Drinkwater, M. Cohen, H. Bogan, E. Joyce, S. Brewster, L. Kunkel, and A. Puca. Life-long sustained mortality advantage of siblings of centenarians. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 99(12):8442–8447, June 2002.
- [476] M. R. S. M. P. A. P.-T. Peter D Karp. The MetaCyc Database. *Nucleic Acids Research*, 30(1):59, Jan. 2002.
- [477] S. Peters, H.-G. Janssen, and G. Vivo-Truyols. Trend analysis of time-series data: A novel method for untargeted metabolite discovery. *Analytica Chimica Acta*, 663(1):98–104, 2010.
- [478] M. Petrascheck, X. Ye, and L. B. Buck. An antidepressant that extends lifespan in adult *Caenorhabditis elegans*. *Nature*, 450(7169):553–556, Nov. 2007.
- [479] G. J. Pickering. Optimizing the sensory characteristics and acceptance of canned cat food: use of a human taste panel. *Proteomics*, 93(1):52–60, Feb. 2009.
- [480] S. B. Pierce, M. Costa, R. Wisotzkey, S. Devadhar, S. A. Homburger, A. R. Buchman, K. C. Ferguson, J. Heller, D. M. Platt, A. A. Pasquinelli, L. X. Liu, S. K. Doberstein, and G. Ruvkun. Regulation of DAF-2 receptor signaling by human insulin and ins-1, a member of the unusually large and diverse *C. elegans* insulin gene family. *Genes & Development*, 15(6):672–686, Mar. 2001.
- [481] M. D. W. Piper, W. Mair, and L. Partridge. Counting the calories: the role of specific nutrients in extension of life span by food restriction. *The Journals of Gerontology Series A, Biological Sciences and Medical Sciences*, 60(5):549–555, May 2005.

- [482] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Oresic. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11:395, 2010.
- [483] C. Pujol, I. Bratic-Hench, M. Sumakovic, J. Hench, A. Mourier, L. Baumann, V. Pavlenko, and A. Trifunovic. Succinate dehydrogenase upregulation destabilize complex I and limits the lifespan of gas-1 mutant. *PloS one*, 8(3):e59493, 2013.
- [484] C. Pungaliya, J. Srinivasan, B. W. Fox, R. U. Malik, A. H. Ludewig, P. W. Sternberg, and F. C. Schroeder. A shortcut to identifying small molecule signals that regulate behavior and development in *Caenorhabditis elegans*. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 106(19):7708–7713, May 2009.
- [485] L. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M. Walsh, J. Berden, K. Brindle, D. Kell, J. Rowland, H. Westerhoff, K. van Dam, and S. Oliver. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19(1):45–50, 2001.
- [486] P. D. Rainville, C. L. Stumpf, J. P. Shockcor, R. S. Plumb, and J. K. Nicholson. Novel application of reversed-phase UPLC-oeTOF-MS for lipid analysis in complex biological mixtures: a new tool for lipidomics. *Journal of Proteome Research*, 6(2):552–558, Feb. 2007.
- [487] S. L. Rea, N. Ventura, and T. E. Johnson. Relationship between mitochondrial electron transport chain dysfunction, development, and life extension in *Caenorhabditis elegans*. *PLoS Biology*, 5(10):e259, Oct. 2007.
- [488] S. N. Reinke, X. Hu, B. D. Sykes, and B. D. Lemire. *Caenorhabditis elegans* diet significantly affects metabolic profile, mitochondrial DNA levels, lifespan and brood size. *Molecular Genetics And Metabolism*, 100(3):274–282, 2010.
- [489] R. J. S. Reis, L. Xu, H. Lee, M. Chae, J. J. Thaden, P. Bharill, C. Tazearslan, E. Siegel, R. Alla, P. Zimniak, and S. Ayyadevara. Modulation of lipid biosynthesis contributes to stress resistance and longevity of *C. elegans* mutants. *Aging*, 3(2):125–147, Feb. 2011.
- [490] J. Requena, C. Chao, R. Levine, and E. Stadtman. Glutamic and aminoadipic semi-aldehydes are the main carbonyl products of metal-catalyzed oxidation of proteins. *Proceedings of the National Academy Of Sciences*, 98(1):69–74, 2001.
- [491] K. Richter, H. Nygren, P. Malmberg, and B. Hagenhoff. Localization of fatty acids with selective chain length by imaging time-of-flight secondary ion mass spectrometry. *Microscopy Research and Technique*, 70(7):640–647, July 2007.

- [492] D. L. Riddle, T. Blumenthal, B. J. Meyer, J. R. Priess, D. L. Riddle, and P. S. Albert. *Genetic and Environmental Regulation of Dauer Larva Development. C. elegans II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2nd edition, 1997.
- [493] M. Ristow and S. Schmeisser. Extending life span by increasing oxidative stress. *Free Radical Biology & Medicine*, 51(2):327–336, July 2011.
- [494] M. Ristow and K. Zarse. How increased oxidative stress promotes longevity and metabolic health: The concept of mitochondrial hormesis (mitohormesis). *Experimental Gerontology*, 45(6):410–418, June 2010.
- [495] G. Rizki, T. N. Iwata, J. Li, C. G. Riedel, C. L. Picard, M. Jan, C. T. Murphy, and S. S. Lee. The Evolutionarily Conserved Longevity Determinants HCF-1 and SIR-2.1/SIRT1 Collaborate to Regulate DAF-16/FOXO. *PLoS genetics*, 7(9):e1002235, Sept. 2011.
- [496] L. D. Roberts, G. McCombie, C. M. Titman, and J. L. Griffin. A matter of fat: an introduction to lipidomic profiling methods. *Journal Of Chromatography B-Analytical Technologies In The Biomedical And Life Sciences*, 871(2):174–181, Aug. 2008.
- [497] S. Robida-Stubbs, K. Glover-Cutter, D. W. Lamming, M. Mizunuma, S. D. Narasimhan, E. Neumann-Haefelin, D. M. Sabatini, and T. K. Blackwell. TOR signaling and rapamycin influence longevity by regulating SKN-1/Nrf and DAF-16/FoxO. *Cell Metabolism*, 15(5):713–724, May 2012.
- [498] S. L. Robinette, K. A. Veselkov, E. Bohus, M. Coen, H. C. Keun, T. M. D. Ebbels, O. Beckonert, E. C. Holmes, J. C. Lindon, and J. K. Nicholson. Cluster analysis statistical spectroscopy using nuclear magnetic resonance generated metabolic data sets from perturbed biological systems. *Analytical Chemistry*, 81(16):6581–6589, Aug. 2009.
- [499] M. D. Robinson, D. P. De Souza, W. W. Keen, E. C. Saunders, M. J. McConville, T. P. Speed, and V. A. Likić. A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC Bioinformatics*, 8:419, 2007.
- [500] D. Rocke and B. Durbin. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics (Oxford, England)*, 19(8):966–972, 2003.
- [501] R. P. Rodgers, E. N. Blumer, C. L. Hendrickson, and A. G. Marshall. Stable isotope incorporation triples the upper mass limit for determination of elemental composition by accurate mass measurement. *Journal Of The American Society For Mass Spectrometry*, 11(10):835–840, Oct. 2000.
- [502] B. Rogina. Longevity Regulation by Drosophila Rpd3 Deacetylase and Caloric Restriction. *Science (New York, NY)*, 298(5599):1745–1745, Nov. 2002.

- [503] C. B. Rohde, F. Zeng, R. Gonzalez-Rubio, M. Angel, and M. F. Yanik. Microfluidic system for on-chip high-throughput whole-animal sorting and screening at subcellular resolution. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 104(35):13891–13895, Aug. 2007.
- [504] E. Rosenberg. The potential of organic (electrospray- and atmospheric pressure chemical ionisation) mass spectrometric techniques coupled to liquid-phase separation for speciation analysis. *Journal Of Chromatography A*, 1000(1-2):841–889, June 2003.
- [505] D. C. Rubinsztein, G. Mariño, and G. Kroemer. Autophagy and Aging. *Cell*, 146(5):682–695, Sept. 2011.
- [506] T. Saito, S. Nakaie, M. Kinoshita, T. Ihara, S. Kinugasa, A. Nomura, and T. Maeda. Practical guide for accurate quantitative solution state NMR analysis. *Metrologia*, 41(3):213–218, Apr. 2004.
- [507] C. Samara, C. B. Rohde, C. L. Gilleland, S. Norton, S. J. Haggarty, and M. F. Yanik. Large-scale in vivo femtosecond laser neurosurgery screen reveals small-molecule enhancer of regeneration. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 107(43):18342–18347, Oct. 2010.
- [508] Y. Sancak, T. R. Peterson, Y. D. Shaul, R. A. Lindquist, C. C. Thoreen, L. Bar-Peled, and D. M. Sabatini. The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science (New York, NY)*, 320(5882):1496–1501, June 2008.
- [509] G. Sanchez. *plsdepot: Partial Least Squares (PLS) Data Analysis Methods*, r package version 0.1.17 edition, Jan. 2012.
- [510] C. J. Sands, M. Coen, T. M. D. Ebbels, E. Holmes, J. C. Lindon, and J. K. Nicholson. Data-driven approach for metabolite relationship recovery in biological ¹H NMR data sets using iterative statistical total correlation spectroscopy. *Analytical Chemistry*, 83(6):2075–2082, Mar. 2011.
- [511] T. Sangster, H. Major, R. Plumb, A. J. Wilson, and I. D. Wilson. A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. *The Analyst*, 131(10):1075–1078, Oct. 2006.
- [512] T. P. Sangster, J. E. Wingate, L. Burton, F. Teichert, and I. D. Wilson. Investigation of analytical variation in metabonomic analysis using liquid chromatography/mass spectrometry. *Rapid Communications In Mass Spectrometry*, 21(18):2965–2970, 2007.
- [513] J. Saric, E. J. Want, U. Duthaler, M. Lewis, J. Keiser, J. P. Shockcor, G. A. Ross, J. K. Nicholson, E. Holmes, and M. F. M. Tavares. Systematic Evaluation of Extraction Methods for Multiplatform-Based Metabotyping: Application to the *Fasciola hepatica* Metabolome. *Analytical Chemistry*, 84(16):6963–6972, Aug. 2012.

- [514] P. Sarup, S. M. M. Pedersen, N. C. Nielsen, A. Malmendal, and V. Loeschcke. The metabolic profile of long-lived *Drosophila melanogaster*. *PloS one*, 7(10):e47461, 2012.
- [515] F. Savorani, G. Tomasi, and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal Of Magnetic Resonance*, 202(2):190–202, 2010.
- [516] J. C. Schafer, M. E. Winkelbauer, C. L. Williams, C. J. Haycraft, R. A. Desmond, and B. K. Yoder. IFTA-2 is a conserved cilia protein involved in pathways regulating longevity and dauer formation in *Caenorhabditis elegans*. *Journal Of Cell Science*, 119(Pt 19):4088–4100, Oct. 2006.
- [517] A. Schlosser and R. Volkmer Engert. Volatile polydimethylcyclsiloxanes in the ambient laboratory air identified as source of extreme background signals in nano-electrospray mass spectrometry. *Journal of Mass Spectrometry*, 38(5):523–525, 2003.
- [518] T. Schulz, K. Zarse, A. Voigt, N. Urban, M. Birringer, and M. Ristow. Glucose restriction extends *Caenorhabditis elegans* life span by inducing mitochondrial respiration and increasing oxidative stress. *Cell Metabolism*, 6:280–293, 2007.
- [519] E. Schuster, J. McElwee, and J. Tullet. DamID in *C. elegans* reveals longevity-associated targets of DAF-16/FoxO. *Molecular Systems Biology*, 6(399), 2010.
- [520] D. Sell, C. Strauch, W. Shen, and V. Monnier. 2-Aminoadipic acid is a marker of protein carbonyl oxidation in the aging human skin: effects of diabetes, renal failure and sepsis. *Biochemical Journal*, 404:269–277, 2007.
- [521] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.
- [522] D. D. Shaye and I. Greenwald. OrthoList: A Compendium of *C. elegans* Genes with Human Orthologs. *PloS one*, 6(5):e20085, May 2011.
- [523] K. L. Sheaffer, D. L. Updike, and S. E. Mango. The Target of Rapamycin pathway antagonizes pha-4/FoxA to control development and aging. *Current Biology*, 18(18):1355–1364, Sept. 2008.
- [524] R. J. Shmookler Reis, P. Bharill, C. Tazearslan, and S. Ayyadevara. Extreme-longevity mutations orchestrate silencing of multiple signaling pathways. *Biochimica et Biophysica Acta*, 1790(10):1075–1083, Oct. 2009.
- [525] A. J. Simpson, L.-H. Tseng, M. J. Simpson, M. Spraul, U. Braumann, W. L. Kingery, B. P. Kelleher, and M. H. B. Hayes. The application of LC-NMR and LC-SPE-NMR to compositional studies of natural organic matter. *The Analyst*, 129(12):1216–1222, Dec. 2004.

- [526] R. Singh, S. Kaushik, Y. Wang, Y. Xiang, I. Novak, M. Komatsu, K. Tanaka, A. M. Cuervo, and M. J. Czaja. Autophagy regulates lipid metabolism. *Nature*, 458(7242):1131–U64, 2009.
- [527] S. P. Singh, M. Niemczyk, L. Zimniak, and P. Zimniak. Fat accumulation in *Caenorhabditis elegans* triggered by the electrophilic lipid peroxidation product 4-hydroxynonenal (4-HNE). *Aging*, 1(1):68–80, 2009.
- [528] P. Sjövall, J. Lausmaa, H. Nygren, L. Carlsson, and P. Malmberg. Imaging of membrane lipids in single cells by imprint-imaging time-of-flight secondary ion mass spectrometry. *Analytical Chemistry*, 75(14):3429–3434, July 2003.
- [529] Smallcombe, Patt, and Keifer. WET Solvent Suppression and Its Applications to LC NMR and High-Resolution NMR Spectroscopy. *Journal of Magnetic Resonance, Series A*, 117(2):9–9, Dec. 1995.
- [530] C. Smith, G. O’Maille, E. Want, C. Qin, S. Trauger, T. Brandon, D. Custodio, R. Abagyan, and G. Siuzdak. METLIN - A metabolite mass spectral database. *Therapeutic Drug Monitoring*, 27(6):747–751, 2005.
- [531] C. A. Smith, E. J. Want, G. O’Maille, R. Abagyan, and G. Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, Feb. 2006.
- [532] E. D. Smith, T. L. Kaeberlein, B. T. Lydum, J. Sager, K. L. Welton, B. K. Kennedy, and M. Kaeberlein. Age- and calorie-independent life span extension from dietary restriction by bacterial deprivation in *Caenorhabditis elegans*. *BMC Developmental Biology*, 8:49, 2008.
- [533] H. Smith, R. Gnanadesikan, and J. B. Hughes. JSTOR: Biometrics, Vol. 18, No. 1 (Mar., 1962), pp. 22-41. *Biometrics*, 1962.
- [534] K. Spagou, H. Tsoukali, N. Raikos, H. Gika, I. D. Wilson, and G. Theodoridis. Hydrophilic interaction chromatography coupled to MS for metabonomic/metabolomic studies. *Proteomics*, 33(6-7):716–727, Mar. 2010.
- [535] M. Spraul, A. S. Freund, R. E. Nast, R. S. Withers, W. E. Maas, and O. Corcoran. Advancing NMR sensitivity for LC-NMR-MS using a cryoflow probe: application to the analysis of acetaminophen metabolites in urine. *Analytical Chemistry*, 75(6):1536–1541, Mar. 2003.
- [536] J. Srinivasan, F. Kaplan, R. Ajredini, C. Zachariah, H. T. Alborn, P. E. A. Teal, R. U. Malik, A. S. Edison, P. W. Sternberg, and F. C. Schroeder. A blend of small molecules regulates both mating and development in *Caenorhabditis elegans*. *Nature*, 454(7208):1115–1118, Aug. 2008.

- [537] E. Stadtman. Role of oxidant species in aging. *Current Medicinal Chemistry*, 11(9):1105–1112, 2004.
- [538] E. Stadtman and R. Levine. Free radical-mediated oxidation of free amino acids and amino acid residues in proteins. *Amino Acids*, 25:207–218, 2003.
- [539] S. Stein. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal Of The American Society For Mass Spectrometry*, 10(8):770–781, 1999.
- [540] S. Stein and D. SCOTT. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal Of The American Society For Mass Spectrometry*, 5(9):859–866, 1994.
- [541] S. E. Stein, V. I. Babushok, R. L. Brown, and P. J. Linstrom. Estimation of Kováts Retention Indices Using Group Contributions. *Analytical Chemistry*, 47(3):975–980, May 2007.
- [542] C. Steinbeck. Recent developments in automated structure elucidation of natural products. *Natural Product Reports*, 21(4):512, 2004.
- [543] M. L. Steinhauser, A. P. Bailey, S. E. Senyo, C. Guillermier, T. S. Perlstein, A. P. Gould, R. T. Lee, and C. P. Lechene. Multi-isotope imaging mass spectrometry quantifies stem cell division and metabolism. *Nature*, 481(7382):516–519, Jan. 2012.
- [544] T. Stiernagle. Maintenance of *C. elegans*. *WormBook*, pages 1–11, 2006.
- [545] M. Stoeckli, P. Chaurand, D. Hallahan, and R. Caprioli. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine*, 7(4):493–496, 2001.
- [546] G. R. Strasser and I. Váradi. Investigations of artifact peaks in sensitive high-performance liquid chromatography methods. *Journal Of Chromatography A*, 869(1-2):85–90, Feb. 2000.
- [547] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher. OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9:163, 2008.
- [548] M. P. Styczynski, J. F. Moxley, L. V. Tong, J. L. Walther, K. L. Jensen, and G. N. Stephanopoulos. Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Analytical Chemistry*, 79(3):966–973, Feb. 2007.
- [549] P. Styles, N. F. Soffe, C. A. Scott, D. A. Crag, F. Row, D. J. White, and P. C. J. White. A high-resolution NMR probe in which the coil and preamplifier are cooled with liquid helium. *International Journal of Pharmaceutics*, 60(3):397–404, Dec. 1984.

- [550] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, and S. Subramaniam. LMSD: LIPID MAPS structure database. *Nucleic Acids Research*, 35(Database issue):D527–32, Jan. 2007.
- [551] L. Sumner, A. Amberg, D. Barrett, M. Beale, R. Beger, C. Daykin, T. Fan, O. Fiehn, R. Goodacre, and J. L. Griffin. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3):211–221, 2007.
- [552] M. V. Sundaram. RTK/Ras/MAPK signaling. *WormBook*, 2005.
- [553] J. R. Swann, E. J. Want, F. M. Geier, K. Spagou, I. D. Wilson, J. E. Sidaway, J. K. Nicholson, and E. Holmes. Systemic gut microbial modulation of bile acid metabolism in host tissue compartments. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 108 Suppl 1:4523–4530, Mar. 2011.
- [554] L. W. Swanson. Liquid metal ion sources: Mechanism and applications. *Nuclear Instruments and Methods in Physics Research*, 218(1-3):347–353, Dec. 1983.
- [555] M. E. Swartz and B. J. Murphy. Ultra performance liquid chromatography: Tomorrow’s HPLC technology today. *LabPlus Int*, 2004.
- [556] J. Swire, S. Fuchs, J. G. Bundy, and A. M. Leroi. The cellular geometry of growth drives the amino acid economy of *Caenorhabditis elegans*. *Proceedings Biological Sciences - The Royal Society*, 276(1668):2747–2754, Aug. 2009.
- [557] P. Syntichaki, K. Troulinaki, and N. Tavernarakis. eIF4E function in somatic cells modulates ageing in *Caenorhabditis elegans*. *Nature*, 445(7130):922–926, Feb. 2007.
- [558] S. S. W. Szeto, S. N. Reinke, and B. D. Lemire. (1)H NMR-based metabolic profiling reveals inherent biological variation in yeast and nematode model systems. *Journal of Biomolecular NMR*, 49(3-4):245–254, Apr. 2011.
- [559] R. Taguchi, T. Houjou, H. Nakanishi, T. Yamazaki, M. Ishida, M. Imagawa, and T. Shimizu. Focused lipidomics by tandem mass spectrometry. *Journal Of Chromatography B-Analytical Technologies In The Biomedical And Life Sciences*, 823(1):26–36, Aug. 2005.
- [560] L. S. Tain, E. Lozano, A. G. Sáez, and A. M. Leroi. Dietary regulation of hypodermal polyploidization in *C. elegans*. *BMC Developmental Biology*, 8(1):28, 2008.
- [561] M. Tan, G. Jones, G. Zhu, J. Ye, T. Hong, S. Zhou, S. Zhang, and L. Zhang. Fellatio by Fruit Bats Prolongs Copulation Time. *PloS one*, 4(10):e7595, Oct. 2009.
- [562] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, and T. Matsuo. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications In Mass Spectrometry*, 2(8):151–153, Aug. 1988.

- [563] T. Tanaka, K. Ikita, T. Ashida, Y. Motoyama, Y. Yamaguchi, and K. Satouchi. Effects of growth temperature on the fatty acid composition of the free-living nematode *Caenorhabditis elegans*. *Lipids*, 31(11):1173–1178, Nov. 1996.
- [564] A. Tannús and M. Garwood. Adiabatic pulses. *Nmr In Biomedicine*, 1997.
- [565] M. Tatar, A. Kopelman, D. Epstein, M. P. Tu, C. M. Yin, and R. S. Garofalo. A mutant *Drosophila* insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science (New York, NY)*, 292(5514):107–110, Apr. 2001.
- [566] R. Tautenhahn, C. Bottcher, and S. Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9:504, 2008.
- [567] R. C. Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, Jan. 2013.
- [568] D. Thierry-Mieg and J. Thierry-Mieg. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biology*, 7 Suppl 1:S12.1–14, 2006.
- [569] J. H. Thomas. Lifespan: The effects of sensory deprivation : Article : Nature. *Nature*, 402(6763):740–741, Dec. 1999.
- [570] Y. Tikunov, A. Lommen, C. H. R. de Vos, H. A. Verhoeven, R. J. Bino, R. D. Hall, and A. G. Bovy. A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiology*, 139(3):1125–1137, Nov. 2005.
- [571] H. Tissenbaum and L. Guarente. Increased dosage of a sir-2 gene extends lifespan in *Caenorhabditis elegans*. *Nature*, 410(6825):227–230, 2001.
- [572] R. t’Kindt, K. Morreel, D. Deforce, W. Boerjan, and J. Van Bocxlaer. Joint GC-MS and LC-MS platforms for comprehensive plant metabolomics: repeatability and sample pre-treatment. *Journal Of Chromatography B-Analytical Technologies In The Biomedical And Life Sciences*, 877(29):3572–3580, Nov. 2009.
- [573] A. Tomás-Loba, B. Bernardes de Jesus, J. M. Mato, and M. A. Blasco. A metabolic signature predicts biological age in mice. *Aging Cell*, 12(1):93–101, Feb. 2013.
- [574] H. Tong, D. Bell, K. Tabei, and M. M. Siegel. Automated data massaging, interpretation, and e-mailing modules for high throughput open access mass spectrometry. *Journal Of The American Society For Mass Spectrometry*, 10(11):1174–1187, 1999.
- [575] K. Tonosaki, Y. Hori, Y. Shimizu, and K. Tonosaki. Relationships between insulin release and taste. *Biomedical Research*, 28(2):79–83, Apr. 2007.

- [576] M. L. Tóth, T. Sigmond, E. Borsos, J. Barna, P. Erdélyi, K. Takacs-Vellai, L. Orosz, A. L. Kovacs, G. Csikós, M. Sass, and T. Vellai. Longevity pathways converge on autophagy genes to regulate life span in *Caenorhabditis elegans*. *Autophagy*, 4(3):330–338, Apr. 2008.
- [577] D. Touboul, F. Kollmer, E. Niehuis, A. Brunelle, and O. Laprévote. Improvement of biological time-of-flight-secondary ion mass spectrometry imaging with a bismuth cluster ion source. *Journal Of The American Society For Mass Spectrometry*, 16(10):1608–1618, Oct. 2005.
- [578] G. D. Tredwell, V. Behrends, F. M. Geier, M. Liebeke, and J. G. Bundy. Between-Person Comparison of Metabolite Fitting for NMR-Based Quantitative Metabolomics. *Analytical Chemistry*, 83(22):8683–8687, Nov. 2011.
- [579] J. Trygg, E. Holmes, and T. Lundstedt. Chemometrics in metabonomics. *Journal of Proteome Research*, 6(2):469–479, Feb. 2007.
- [580] D. D. Tudor, C. Ahia, and D. W. Montgomerie. Controlling evaporation lessens ethanol concentration variability between NGM plates. *The Worm Breeder's Gazette*, 18(1), Dec. 2009.
- [581] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley series in behavioral sciences. Addison-Wesley Publishing Company, 1977.
- [582] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy Of Sciences*, 98(9):5116–5121, Apr. 2001.
- [583] M. Tyers and M. Mann. From genomics to proteomics. *Nature*, 422(6928):193–197, Mar. 2003.
- [584] K. Uchida. Histidine and lysine as targets of oxidative modification. *Amino Acids*, 25(3):249–257, 2003.
- [585] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Research*, 36(Database issue):D402–8, Jan. 2008.
- [586] J. J. Van Deemter, F. J. Zuiderweg, and A. Klinkenberg. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. *Chemical Engineering Science*, 50(24):3869–3882, 1956.
- [587] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7:142, 2006.

- [588] M. J. van der Werf, K. M. Overkamp, B. Muilwijk, L. Coulier, and T. Hankemeier. Microbial metabolomics: Toward a platform with full metabolome coverage. *Analytical Biochemistry*, 370(1):17–25, 2007.
- [589] J. M. Van Raamsdonk and S. Hekimi. Deletion of the mitochondrial superoxide dismutase sod-2 extends lifespan in *Caenorhabditis elegans*. *PLoS genetics*, 5(2):e1000361, Feb. 2009.
- [590] J. M. Van Raamsdonk and S. Hekimi. FUDR causes a twofold increase in the lifespan of the mitochondrial mutant gas-1. *Mechanisms Of Ageing And Development*, 132(10):519–521, Oct. 2011.
- [591] J. M. Van Raamsdonk and S. Hekimi. Superoxide dismutase is dispensable for normal animal lifespan. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 109(15):5785–5790, Apr. 2012.
- [592] W. A. Van Voorhies. Production of sperm reduces nematode lifespan. *Nature*, 360(6403):456–458, Dec. 1992.
- [593] J. Vaupel and K. v. Kistowski. Broken limits to life expectancy. *Life*, 2005.
- [594] J. W. Vaupel. Biodemography of human ageing. *Nature*, 464(7288):536–542, 2010.
- [595] J. vB Hjelmberg, I. Iachine, A. Skyttje, J. W. Vaupel, M. McGue, M. Koskenvuo, J. Kaprio, N. L. Pedersen, and K. Christensen. Genetic influence on human lifespan and longevity. *Human Genetics*, 119(3):312–321, Apr. 2006.
- [596] T. Vellai, K. Takacs-Vellai, Y. Zhang, A. L. Kovacs, L. Orosz, and F. Müller. Genetics: influence of TOR kinase on lifespan in *C. elegans*. *Nature*, 426(6967):620, Dec. 2003.
- [597] K. M. Verge and G. R. Agnes. Plasticizer contamination from vacuum system O-rings in a quadrupole ion trap mass spectrometer. *Journal Of The American Society For Mass Spectrometry*, 13(8):901–905, 2002.
- [598] K. A. Veselkov, J. C. Lindon, T. M. D. Ebbels, D. Crockford, V. V. Volynkin, E. Holmes, D. B. Davies, and J. K. Nicholson. Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery. *Analytical Chemistry*, 81(1):56–66, Jan. 2009.
- [599] K. A. Veselkov, L. K. Vingara, P. Masson, S. L. Robinette, E. Want, J. V. Li, R. H. Barton, C. Boursier-Neyret, B. Walther, T. M. Ebbels, I. Pelczer, E. Holmes, J. C. Lindon, and J. K. Nicholson. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry*, 83(15):5864–5872, Aug. 2011.

- [600] J. Vessman, R. I. Stefan, J. F. Van Staden, K. Danzer, W. Lindner, D. T. Burns, A. Fajgelj, and H. Muller. Selectivity in analytical chemistry. *Pure Appl. Chem*, 73(8):1381–1386, 2001.
- [601] B. Virk, G. Correia, D. P. Dixon, I. Feyst, J. Jia, N. Oberleitner, Z. Briggs, E. Hodge, R. Edwards, J. Ward, D. Gems, and D. Weinkove. Excessive folate synthesis limits lifespan in the *C. elegans*: E. coli aging model. *BMC Biology*, 10(1):67, 2012.
- [602] M. Viswanathan and L. Guarente. Regulation of *Caenorhabditis elegans* lifespan by sir-2.1 transgenes. *Nature*, 477(7365):E1–E2, Sept. 2011.
- [603] S. H. von Reuss, N. Bose, J. Srinivasan, J. J. Yim, J. C. Judkins, P. W. Sternberg, and F. C. Schroeder. Comparative Metabolomics Reveals Biogenesis of Ascarosides, a Modular Library of Small-Molecule Signals in *C. elegans*. *Journal of the American Chemical Society*, 134(3):1817–1824, Jan. 2012.
- [604] O. Vorst, C. H. R. d. Vos, A. Lommen, R. V. Staps, R. G. F. Visser, R. J. Bino, and R. D. Hall. A non-directed approach to the differential analysis of multiple LC–MS-derived metabolic profiles. *Metabolomics*, 1(2):169–180, Apr. 2005.
- [605] G. Vuister and A. Bax. Resolution enhancement and spectral editing of uniformly ^{13}C -enriched proteins by homonuclear broadband ^{13}C decoupling. *Journal Of Magnetic Resonance*, 1992.
- [606] G. S. Walker and T. N. O’Connell. Comparison of LC-NMR and conventional NMR for structure elucidation in drug metabolism studies. *Expert Opin. Drug Metab. Toxicol.*, 4(10):1295–1305, 2008.
- [607] M. C. Wang, D. Bohmann, and H. Jasper. JNK extends life span and limits growth by antagonizing cellular and organism-wide responses to insulin signaling. *Cell*, 121(1):115–125, Apr. 2005.
- [608] M. C. Wang, E. J. O’Rourke, and G. Ruvkun. Fat metabolism links germline stem cells and longevity in *C. elegans*. *Science (New York, NY)*, 322(5903):957–960, Nov. 2008.
- [609] M. C. Wang, E. J. O’Rourke, and G. Ruvkun. Fat metabolism links germline stem cells and longevity in *C. elegans*. *Science (New York, NY)*, 322(5903):957–960, Nov. 2008.
- [610] S.-C. Wang, C.-M. Huang, and S.-M. Chiang. Improving signal-to-noise ratios of liquid chromatography-tandem mass spectrometry peaks using noise frequency spectrum modification between two consecutive matched-filtering procedures. *Journal Of Chromatography A*, 1161(1-2):192–197, Aug. 2007.

- [611] Y. Wang, D. Lawler, B. Larson, Z. Ramadan, S. Kochhar, E. Holmes, and J. K. Nicholson. Metabonomic investigations of aging and caloric restriction in a life-long dog study. *Journal of Proteome Research*, 6(5):1846–1854, 2007.
- [612] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, Jan. 2009.
- [613] E. J. Want, P. Masson, F. Michopoulos, I. D. Wilson, G. Theodoridis, R. S. Plumb, J. Shockcor, N. Loftus, E. Holmes, and J. K. Nicholson. Global metabolic profiling of animal and human tissues via UPLC-MS. *Nature Protocols*, 8(1):17–32, Dec. 2012.
- [614] E. J. Want, A. Nordstrom, H. Morita, and G. Siuzdak. From exogenous to endogenous: the inevitable imprint of mass spectrometry in metabolomics. *Journal of Proteome Research*, 6(2):459–468, 2007.
- [615] E. J. Want, G. O’Maille, C. A. Smith, T. R. Brandon, W. Uritboonthai, C. Qin, S. A. Trauger, and G. Siuzdak. Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Analytical Chemistry*, 78(3):743–752, Feb. 2006.
- [616] E. J. Want, I. D. Wilson, H. Gika, G. Theodoridis, R. S. Plumb, J. Shockcor, E. Holmes, and J. K. Nicholson. Global metabolic profiling procedures for urine using UPLC-MS. *Nature Protocols*, 5(6):1005–1018, 2010.
- [617] A. M. Weljie, J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky. Targeted profiling: quantitative analysis of ¹H NMR metabolomics data. *Analytical Chemistry*, 78(13):4430–4442, July 2006.
- [618] P. Wenig and J. Odermatt. OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. *BMC Bioinformatics*, 11:405, 2010.
- [619] E. Werner, J.-F. Heilier, C. Ducruix, E. Ezan, C. Junot, and J.-C. Tabet. Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends. *Journal Of Chromatography B-Analytical Technologies In The Biomedical And Life Sciences*, 871(2):143–163, Aug. 2008.
- [620] R. G. Westendorp and T. B. Kirkwood. Human longevity at the cost of reproductive success. *Nature*, 396(6713):743–746, 1998.
- [621] H. V. Westerhoff and B. O. Palsson. The evolution of molecular biology into systems biology. *Nature Biotechnology*, 22(10):1249–1252, Oct. 2004.
- [622] J. A. Westerhuis, H. C. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. van Velzen, J. P. van Duijnhoven, and F. A. van Dorsten. Assessment of PLS-DA cross validation. *Metabolomics*, 4(1):81–89, 2008.

- [623] G. M. Whitesides. The origins and the future of microfluidics. *Nature*, 442(7101):368–373, July 2006.
- [624] M. E. Wieser. Atomic weights of the elements 2005 (IUPAC Technical Report). *Pure Appl. Chem*, 78(11):2051–2066, 2006.
- [625] A. Wijeyesekera, C. Selman, R. H. Barton, E. Holmes, J. K. Nicholson, and D. J. Withers. Metabotyping of long-lived mice using ^1H NMR spectroscopy. *Journal of Proteome Research*, 11(4):2224–2235, Apr. 2012.
- [626] D. S. Williams, A. Cash, L. Hamadani, and T. Diemer. Oxaloacetate supplementation increases lifespan in *Caenorhabditis elegans* through an AMPK/FOXO-dependent pathway. *Aging Cell*, 8(6):765–768, 2009.
- [627] G. Williams. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution*, 11(4):398–411, Dec. 1957.
- [628] R. Williams, E. Lenz, A. Wilson, J. Granger, I. Wilson, H. Major, C. Stumpf, and R. Plumb. A multi-analytical platform approach to the metabonomic analysis of plasma from normal and zucker (fa/fa) obese rats. *Molecular Biosystems*, 2(3-4):174–183, 2006.
- [629] S. Williams. Ghost peaks in reversed-phase gradient HPLC: a review and update. *Journal Of Chromatography A*, 1052(1-2):1–11, Oct. 2004.
- [630] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser. HMDB: the human metabolome database. *Nucleic Acids Research*, 35:D521–D526, 2007.
- [631] M. Wolf, F. Nunes, A. Henkel, A. Heinick, and R. J. Paul. The MAP kinase JNK-1 of *Caenorhabditis elegans*: Location, activation, and influences over temperature-dependent insulin-like signaling, stress responses, and fitness. *Journal of Cellular Physiology*, 214(3):721–729, 2008.
- [632] S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11:148, 2010.
- [633] M. M. Wolff and W. E. Stephens. A pulsed mass spectrometer with time dispersion. *Review of Scientific Instruments*, 24:616, 1953.
- [634] S. Wolff and A. Dillin. The trifecta of aging in *Caenorhabditis elegans*. *Experimental Gerontology*, 41(10):894–903, Oct. 2006.

- [635] C. A. Wolkow, M. J. Muñoz, D. L. Riddle, and G. Ruvkun. Insulin Receptor Substrate and p55 Orthologous Adaptor Proteins Function in the *Caenorhabditis elegans* daf-2/Insulin-like Signaling Pathway. *Journal Of Biological Chemistry*, 277(51):49591–49597, 2002.
- [636] A. Wong, P. Boutis, and S. Hekimi. Mutations in the clk-1 gene of *Caenorhabditis elegans* affect developmental and behavioral timing. *Genetics*, 139(3):1247–1259, Mar. 1995.
- [637] D. K. Woo and G. S. Shadel. Mitochondrial Stress Signals Revise an Old Aging Theory. *Cell*, 144(1):11–12, Jan. 2011.
- [638] B. Worley, S. Halouska, and R. Powers. Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical Biochemistry*, 433(2):102–104, Feb. 2013.
- [639] L. Wu, M. R. Mashego, J. C. van Dam, A. M. Proell, J. L. Vinke, C. Ras, W. A. van Winden, W. M. van Gulik, and J. J. Heijnen. Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly ¹³C-labeled cell extracts as internal standards. *Analytical Biochemistry*, 336(2):164–171, Jan. 2005.
- [640] Q.-L. WU, Q. Rui, K.-W. HE, L.-L. SHEN, and D.-Y. WANG. UNC-64 and RIC-4, the plasma membrane-associated SNAREs syntaxin and SNAP-25, regulate fat storage in nematode *Caenorhabditis elegans*. *Neuroscience Bulletin*, 26(2):104–116, Apr. 2010.
- [641] Z. Wu, W. Gao, M. A. Phelps, D. Wu, D. D. Miller, and J. T. Dalton. Favorable Effects of Weak Acids on Negative-Ion Electrospray Ionization Mass Spectrometry. *Analytical Chemistry*, 76(3):839–847, Feb. 2004.
- [642] M. P. Wymann and R. Schneter. Lipid signalling in disease. *Nature Reviews Molecular Cell Biology*, 9(2):162–176, Feb. 2008.
- [643] Y. Xi, J. S. de Ropp, M. R. Viant, D. L. Woodruff, and P. Yu. Improved identification of metabolites in complex mixtures using HSQC NMR spectroscopy. *Analytica Chimica Acta*, 614(2):127–133, May 2008.
- [644] Y. Xi and D. M. Rocke. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*, 9:324, 2008.
- [645] Y.-Q. Xia, S. Patel, R. Bakhtiar, R. B. Franklin, and G. A. Doss. Identification of a new source of interference leached from polypropylene tubes in mass-selective analysis. *Chromatographia*, 16(3):417–421, Mar. 2005.
- [646] W. Xu, H. Yang, Y. Liu, Y. Yang, P. Wang, S.-H. Kim, S. Ito, C. Yang, P. Wang, M.-T. Xiao, L.-x. Liu, W.-q. Jiang, J. Liu, J.-y. Zhang, B. Wang, S. Frye, Y. Zhang,

- Y.-h. Xu, Q.-y. Lei, K.-L. Guan, S.-m. Zhao, and Y. Xiong. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of Î²-ketoglutarate-dependent dioxygenases. *Cancer Cell*, 19(1):17–30, Jan. 2011.
- [647] T. M. Yamawaki, J. R. Berman, M. Suchanek-Kavipurapu, M. McCormick, M. Maria Gaglia, S.-J. Lee, and C. Kenyon. The somatic reproductive tissues of *C. elegans* promote longevity through steroid hormone signaling. *PLoS Biology*, 8(8), 2010.
- [648] W. Yang and S. Hekimi. A mitochondrial superoxide signal triggers increased longevity in *Caenorhabditis elegans*. *PLoS Biology*, 8(12):e1000556, 2010.
- [649] W. Yang and S. Hekimi. Two modes of mitochondrial dysfunction lead independently to lifespan extension in *Caenorhabditis elegans*. *Aging Cell*, 9(3):433–447, June 2010.
- [650] Z. Yang and D. J. Klionsky. Eaten alive: a history of macroautophagy. *Nature Cell Biology*, 12(9):814–822, Sept. 2010.
- [651] M. E. Yanos, C. F. Bennett, and M. Kaeberlein. Genome-Wide RNAi Longevity Screens in *Caenorhabditis elegans*. *Current Genomics*, 13(7):508–518, Nov. 2012.
- [652] D. Ye, S. Ma, Y. Xiong, and K.-L. Guan. R-2-hydroxyglutarate as the key effector of IDH mutations promoting oncogenesis. *Cancer Cell*, 23(3):274–276, Mar. 2013.
- [653] K. Yen, S. D. Narasimhan, and H. A. Tissenbaum. DAF-16/Forkhead box O transcription factor: many paths to a single Fork(head) in the road. *Antioxidants & Redox Signaling*, 14(4):623–634, Feb. 2011.
- [654] K. Yook, T. W. Harris, T. Bieri, A. Cabunoc, J. Chan, W. J. Chen, P. Davis, N. de la Cruz, A. Duong, R. Fang, U. Ganesan, C. Grove, K. Howe, S. Kadam, R. Kishore, R. Lee, Y. Li, H.-M. Muller, C. Nakamura, B. Nash, P. Ozersky, M. Paulini, D. Raciti, A. Rangarajan, G. Schindelman, X. Shi, E. M. Schwarz, M. Ann Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, J. Hodgkin, M. Berriman, R. Durbin, P. Kersey, J. Spieth, L. Stein, and P. W. Sternberg. WormBase 2012: more genomes, more data, new website. *Nucleic Acids Research*, 40(Database issue):D735–41, Jan. 2012.
- [655] R. Yoshida, T. Tamura, C. Takaoka, K. Harada, A. Kobayashi, Y. Mukai, and E. Fukusaki. Metabolomics-based systematic prediction of yeast lifespan and its application for semi-rational screening of ageing-related mutants. *Aging Cell*, 9(4):616–625, Aug. 2010.
- [656] R. A. Yost and C. G. Enke. Selected ion fragmentation with a tandem quadrupole mass spectrometer. *Journal of the American Chemical Society*, 100(7):2274–2275, 1978.

- [657] Z. Yu, G. Zhai, P. Singmann, Y. He, T. Xu, C. Prehn, W. Römisch-Margl, E. Lattka, C. Gieger, N. Soranzo, J. Heinrich, M. Standl, E. Thiering, K. Mittelstraß, H.-E. Wichmann, A. Peters, K. Suhre, Y. Li, J. Adamski, T. D. Spector, T. Illig, and R. Wang-Sattler. Human serum metabolic profiles are age dependent. *Aging Cell*, 11(6):960–967, Dec. 2012.
- [658] H. Yuan and R. Marmorstein. Red Wine, Toast of the Town (Again). *Science (New York, NY)*, 339(6124):1156–1157, 2013.
- [659] K. Zahradka, D. Slade, A. Bailone, S. Sommer, D. Aeverbeck, M. Petranovic, A. B. Lindner, and M. Radman. Reassembly of shattered chromosomes in *Deinococcus radiodurans*. *Nature*, 443(7111):569–573, Oct. 2006.
- [660] S. Zeitoun-Ghandour, O. I. Leszczyszyn, C. A. Blindauer, F. M. Geier, J. G. Bundy, and S. R. Stürzenbaum. *C. elegans* metallothioneins: response to and defence against ROS toxicity. *Molecular Biosystems*, 7(8):2397–2406, June 2011.
- [661] E. Zelena, W. B. Dunn, D. Broadhurst, S. Francis-McIntyre, K. M. Carroll, P. Begley, S. O’Hagan, J. D. Knowles, A. Halsall, H. Consortium, I. D. Wilson, and D. B. Kell. Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Analytical Chemistry*, 81(4):1357–1364, Feb. 2009.
- [662] Y. Zeng, L. Cheng, H. Chen, H. Cao, E. R. Hauser, Y. Liu, Z. Xiao, Q. Tan, X.-L. Tian, and J. W. Vaupel. Effects of FOXO Genotypes on Longevity: A Biodemographic Analysis. *The Journals of Gerontology Series A, Biological Sciences and Medical Sciences*, 65(12):1285–1299, Sept. 2010.
- [663] X. K. Zhang, R. C. Dutky, and H. M. Fales. Rubber Stoppers as Sources of Contaminants in Electrospray Analysis of Peptides and Proteins. *Analytical Chemistry*, 68(18):3288–3289, Jan. 1996.
- [664] Y. Zhang, Z. Shao, Z. Zhai, C. Shen, and J. A. Powell-Coffman. The HIF-1 hypoxia-inducible factor modulates lifespan in *C. elegans*. *PloS one*, 4(7):e6348, 2009.
- [665] H. Zheng, P. C. Ojha, S. McClean, N. D. Black, J. G. Hughes, and C. Shaw. Heuristic charge assignment for deconvolution of electrospray ionization mass spectra. *Rapid Communications in Mass Spectrometry*, 17(5):429–436, 2003.
- [666] L. Zheng, R. T’Kind, S. Decuyper, S. J. von Freyend, G. H. Coombs, and D. G. Watson. Profiling of lipids in *Leishmania donovani* using hydrophilic interaction chromatography in combination with Fourier transform mass spectrometry. *Rapid Communications In Mass Spectrometry*, 24(14):2074–2082, June 2010.

Appendices

Theory and methods

This chapter aims to revise and summarise the basic concepts behind the science[173, 561, 479] and methods used in this thesis and may be referred to, when trying to understand the exact methodologies of the experimental chapters. Particularly as methods are a major emphasis of this work. Even though this chapter is very detailed and concepts are supported by illustrations, whenever possible, it is far from comprehensive. Therefore, the reader is advised to refer to the cited literature.

A1. The basic constituents of matter

All matter is composed of atoms [33, 412], mostly in the form of molecules. Atoms themselves are made up of an atomic nucleus and the negatively charged electrons surrounding it. The nuclei itself is a mixture of the positively charged protons and the non-charge carrying neutrons. A charged atom, *i.e.* where the proton and electron count differ, is also called ion. The charge is written as superscript behind the elemental symbol, such as H^+ for a proton. The number of protons is expressed by its proton number. Each proton number corresponds to a different element, which is written as subscript in front of its elemental symbol, such as ${}_1H$, ${}_7N$ or ${}_{26}Fe$. The proton number defines an element's place in the periodic table and therefore its unique properties. An element, however, with its constant proton count, may exist with several different numbers of neutrons. Those variations of an element are called its isotopes, denoted by the neutron count written as superscript in front of the elemental symbol, such as ${}^1H$ for the protium and ${}^2H$ for the deuterium and ${}^3H$ for the tritium isotope of hydrogen. Isotopes can be stable or subject to radioactive decay with an associated half-life time. In nature many element are present as several isotopes, distributed at a certain ratio. For example, on earth carbon is present as ${}^{12}C$ and ${}^{13}C$, in a ratio of 98.9% to 1.1%. (${}^{14}C$ is only available in trace amounts). This isotope ratio, is also referred to as the natural abundance of an element. When referring

to the molar weight of a molecule, it is important to know, whether it is the mono isotopic weight, *i.e.* the molecule just consists of one specific isotope, or the average molecular weight, which is a weighted average, according to the natural abundance. For example, the average weight of glucose at natural abundance is 180.1559 g/mol , whereas the one of ^{12}C glucose is 180.0633 g/mol . [24]

A2. Analytical chemistry

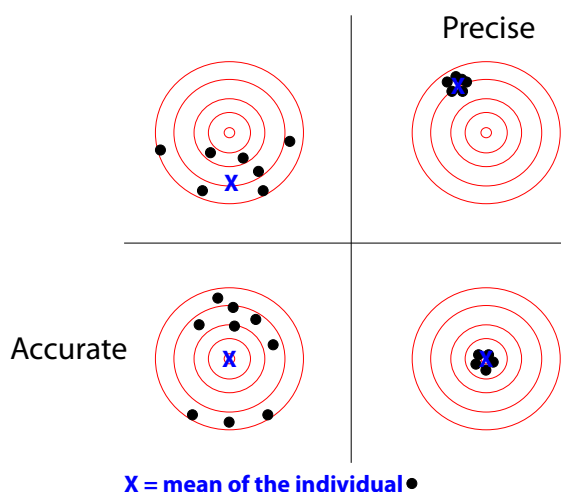


Figure A2.1.: **Precision and accuracy:**

Quantifying a compound in a matrix may be compared to hitting a target.

Analytical chemistry is divided into separation science, qualitative analysis and quantitative analysis. Whereas separation science deals with the separation of mixtures with more than one analyte, the other two refer to the mode of detection. Qualitative analysis is concerned with identifying the analyte, quantitative analysis with determining its concentration within a matrix. Most detection methods of modern analytical chemistry, such as mass spectrometry (MS) or nuclear resonance (NMR) spectroscopy can perform both, qualitative and quantitative analysis. However, they usually have a different concentration threshold. The limit of detection (LOD) is the concentration that is required to be certain about the presence of the specific analyte in the sample. It usually is lower than the limit of quantification

(LOQ); the minimum concentration that can be determined reliably with the given method. Both limits, which are very specific for the analytic method, instrument setup and analyte, depend on the signal to noise ratio (SNR). Albeit there are several ways to express the SNR, they all set the analyte signal in reference to the background signal level. The background usually being a mixture of co-detected undesired chemical entities of the matrix and the electronic noise from the detector. [231]

Additional parameters besides LOD and LOQ, which characterise an analytical method are accuracy, precision, selectivity and robustness. Accuracy measures the proximity of a measurement to the actual (true) value and precision the reproducibility of a measurement, *i.e.* in how far repeated measurements vary. The exact relationship is illustrated in figure A2.1. Selectivity describes in how far a particular analyte inside a matrix or complex mixture can be determined, without any other component's interference [600]. A method's robustness measures its ability to remain unaffected by small deliberate changes to any of its parameters. It is used to characterise its reliability during normal use.

Metabolomics has high requirements for all of these parameters, as it tries to reliably measure as many as possible components of very complex mixtures in a high number of samples. Therefore method development and characterisation should be an integral part of any major metabolomics study.

For more detailed information on definitions and explanations, please refer to the International Union of Pure and Applied Chemistry's (IUPAC) Compendium of Chemical Terminology ("The Gold Book") [405].

A2.1. Separation science

Separation science encompasses all the different methods that are used to convert a complex mixture into two or more products, in order to reduce its complexity prior to detection or isolate and purify any of its components. Those range from crude separation principles, like sieving, distillation, filtration, crystallisation or centrifugation to sophisticated methods such as modern electrophoresis or chromatography. Column chromatography is the only separation method used in this thesis. Its principles will be outlined in this section.

A2.1.1. Basic principles of chromatography

All types of chromatography rely on the interaction of analytes, dissolved in a mobile phase, with a stationary phase. As mobile phase, liquid and gas are used for liquid chromatography and gas chromatography, respectively. The stationary phase is usually bonded with or packed with small particles bonded with functional groups that determine the selectivity of the interaction with the analytes. The more interaction with the stationary phase, the

higher an analyte's retention and therefore the time required to elute. This time, which is very specific (and reproducible) for each analyte on a given system is defined as its retention time (RT). If an analyte is not retained whatsoever, it elutes at the solvent front, also referred to dead time, as it is the minimum time for any molecule required to travel through the void volume of the chromatographic system.

However, retention alone does not yet guarantee a successful chromatographic separation. To separate compounds of very similar retention behaviour, a certain resolving power is required. Resolving power is measured in the number of theoretical plates, a concept loaned from distillation columns. The number of theoretical plates and therefore resolving power is directly proportional to the column length. However, column length is usually limited by the increase in pressure, particularly as modern columns are packed with small functionalised particles. The exact relationship between all these parameters can be modelled using an equation A1, introduced by VanDeemter *et al.*[586].

$$HEPT = A + \frac{A}{u} + C \times u \quad (\text{A1})$$

The resolving power of a column, the height equivalent to a theoretical plate (HEPPT), depends on three terms. A , the Eddy-diffusion, which is diffusion that occurs due to non-ideal packing of the column. The second term is the fraction between a coefficient B , describing the diffusion of the analytes in longitudinal direction and the linear velocity of the mobile phase u . This means the higher the flow rate, the lower the impact of diffusion. The third term is the product between the linear velocity and C , the resistance to mass transfer coefficient. As B and C are fixed, a HEPT versus u curve can be calculated to find the optimal flow rate for a given chromatographic setup.

A2.1.2. Liquid chromatography

A2.1.2.1. Instrumentation

The C-term of the VanDeemter curve depends on the particle size the column is packed with. The smaller the particles, the better the separation efficiency (HEPT). However, the smaller the particles, the higher the back pressure of the system. For the $5\ \mu\text{m}$ diameter particle size of a modern high performance liquid chromatography (HPLC) system, hydrostatic pressure is not sufficient. A pump, generating pressures up to 400 bar is used to

perform the separation. Figure A2.2 illustrates the basic assembly of such a HPLC system.

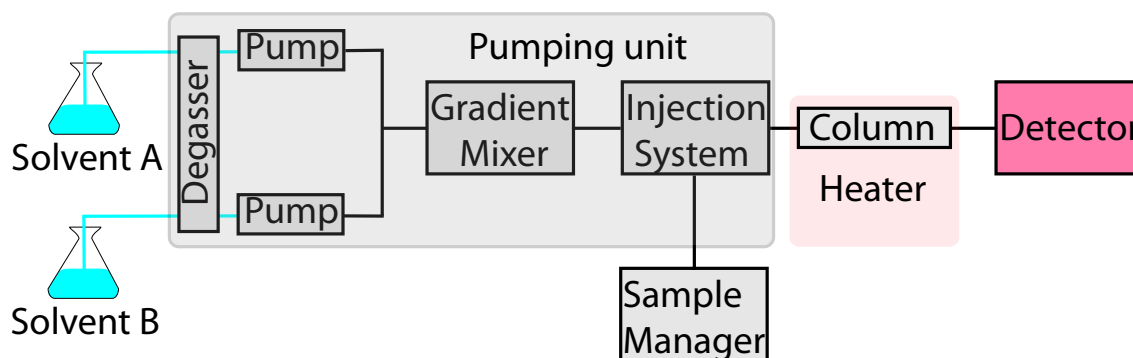


Figure A2.2.: **Liquid chromatography (LC) system:**

Assembly of a typical binary (two different solvents) LC system, such as used for experiments in this thesis.

First, the solvent is degassed to avoid bubbles interfering with the separation. The pumps deliver solvent A and B to the gradient mixer in the user defined ratio and build up the system's operating pressure. The injection system contains an injection loop with defined volume. During an injection, the loop is (fully) filled with the sample, delivered by the needle of a sample manager. Then the solvent flow is diverted through the injection loop and the sample carried onto the column inside the solvent stream. On the column, which is heated to a defined temperature, separation is performed prior to detection (by UV, MS, NMR or a combination thereof).

Recently, a new generation of LC systems has started to replace HPLC systems. By further decreasing the particle size down to $< 2 \mu m$ diameter, ultra performance liquid chromatography (UPLC) systems are able to significantly improve chromatographic resolution and therefore LOD/LOQ and analysis times [555]. According to the vanDeemter curve, this requires a further increase in pressure. Indeed, UPLC system, such as the one that will be used for the LC-MS experiments in this thesis, operate at pressures up to 1000 bar. [101]

A2.1.2.2. Common modes of liquid chromatography

Depending on the column's packing material (stationary phase), the mobile phase and vari-

ous modifiers, many different types of chromatography, each based on a different separation principle and therefore selectivity, can be performed with a single LC setup. Reversed phase (RP), hydrophobic liquid interaction (HILIC) and ion pairing (IP), are the most common chromatography modes for metabolite profiling. As they all are used within this thesis, they will be discussed briefly. However, other important types of chromatography include normal phase, ion exchange, affinity and size exclusion chromatography. [373]

Separations can be isocratic, *i.e.* with no change in the mobile phase composition, or a changing gradient of two (binary) or more mobile phases may be used. The latter makes particular sense if throughout one analysis analytes with highly diverging affinities to the stationary phase have to be separated. The solvent with the weakest eluting power for the given stationary phase is gradually exchanged for a stronger one. If all solvents are ordered by their elution strength, they are arranged in the elutropic series.

Reversed phase chromatography separates analytes based on their lipophilicity. The hydrophobic stationary phase is usually composed of alkanes bonded onto silica beads, with C_{18} being the most prominent chain length. The elutropic series usually ranges from polar solvents like acetic acid, pyridine, water over alcohols like methanol, ethanol to very apolar solvents, such as chloroform, with hexane being the strongest. However, for the majority of all reversed phase applications a gradient from water to methanol or acetonitrile is used. Therefore apolar analytes, such as lipids are well retained, but polar ones, such as small organic acids or sugars elute at the very beginning of a run. It is therefore well suited for the separation of semipolar and apolar molecules, such as encompassed in blood, tissue extracts or urine. Another benefit of this selectivity is that inorganic salts, such as NaCl elute with the void volume and not cause undesired interferences if mass spectrometry is used for detection. [613]

HILIC has been developed by Andrew Alpert [12] to allow the analysis very polar molecules. The selectivity of HILIC is roughly inverse to RP, with the most polar molecules being retained the longest. Its separation mechanism is not fully understood yet, but is likely to be a combination of different mobile and stationary phase interactions, with liquid/liquid partitioning being the most prominent [237]. Typical stationary phases packing for HILIC include plain silica (SiO_2) beads, or functionalisation with aminopropyl (NH_2), cyanopropyl (CN), carbamoyl (Amide) or even some zwitter-ionic groups [295]. In contrast to a re-

versed phase system, HILIC starts at high organic solvent (A) and gradually changes to aqueous conditions (B). Therefore the sample requires to be reconstituted in at least 50:50 water:organic, unlike for reversed phase, where only water is used. Solvent A is usually a > 90% acetonitrile buffer and B an aqueous buffer. The most common buffer systems are ammonium formate or ammonium acetate. HILIC is also well suited for LC-MS as the high organic at start allows for good electrospray performance [29, 534]. [237]

Ion pair chromatography is based on a reversed phase system, but allows for the separation of (very polar,) charged molecules. Additionally to a classical reversed phase method, the solvents are supplemented with an ion pair reagent. Those are amphiphilic molecules, such as tetrabutylamine for anionic separations or heptafluorobutyric acid for cationic separations. The hydrophobic chain(s) are to retain the molecule on a C_{18} phase, the ionic group to interact with the charge of the (oppositely) charged analyte. After the neutral charge complex has bonded to the stationary phase, it can be eluted like a regular reversed phase retained analyte. However, as the retention of the complex is due to the ion pair's hydrophobic group, rather than the compound itself, those complexes usually have very similar retention characteristics. Therefore a gradient of several isocratic regions, specifically tailored for the used ion pair, is preferred over a linear reversed phase gradient. Using ion pair chromatography and mass spectrometric detection can yield very powerful methods [378, 68, 375]. However, it is important to ensure that sensitivity is not compromised by ion suppression by the desired ion pair [217]. [87].

A2.1.3. Gas chromatography

Gas-chromatography is a very established and efficient separation technique [278], suited for small molecules that are volatile or can be made volatile by means of derivatisation. The mobile phase is an inert carrier gas, such as helium or hydrogen. As stationary phase either packed or capillary columns are used. In the second case the stationary phase is a liquid, absorbed to a carrier material. As capillary columns achieve much higher chromatographic resolution, they will be used for experiments in this thesis.

A2.1.3.1. Instrumentation

Figure A2.3 outlines the principal assembly of a gas chromatograph.

A defined amount of the sample is injected into an inert glass liner by the injection syringe

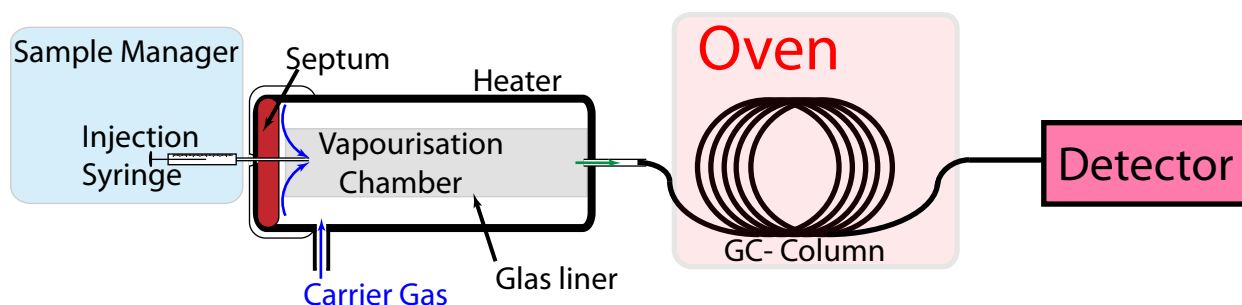


Figure A2.3.: **Gas chromatography (GC) system:**

Assembly of a typical gas chromatography system, such as used for experiments in this thesis.

of the sample manager. A septum ensures air-tightness after the needle is withdrawn. As the glass liner is heated, the sample is evaporated instantly. The steady stream of carrier gas sweeps the volatile sample into the capillary column, where analytes start to interact with the stationary phase.

For experiments in this thesis, a 30 m long, 0.25 μm inner diameter, silicon oil film coated (95% dimethyl/ 5% diphenyl polysiloxane) capillary column was used as stationary phase. On this column, molecules are retained based on their polarity; the less polar, the higher the retention time. In contrast to (RP) LC, where elution is performed by changing the mobile phase polarity, in GC a temperature gradient, controlled by the programmable GC-oven is used for elution. A higher temperature will lead to less interaction between analytes and stationary phase and therefore a shorter retention time. Finally, once a compound is eluted it can be detected. In this thesis, a mass spectrometer was used as detector, as described in section A2.3.3.

A2.1.3.2. Derivatisation

In order to be passed onto the column and separated, GC requires an analyte to be volatile. Some small molecules don't fulfil this prerequisite, due the polarity of their functional groups. However, by converting polar functional groups into less polar ones, derivatisation can make those molecules GC compatible. In this thesis, a two-step derivatisation procedure, consisting of methoximation followed by silylation, is used, as it was shown to be reproducible and have a broad coverage [171, 214, 67]. In the first, the

methoximation step, carbonyls are converted into a Schiff base, by methoxyamine. This prevents sugars from cyclisation (and forces them into open-ring formation) and stabilises α -ketoacids. In the second step, in all functional groups with acidic protons, such as alcohols, amines, carboxyls and thiols, the acidic proton is replaced by a silyl group. Two silylation reagents are particularly common in metabolomic studies. MSTFA (N-methyl-trimethylsilyltrifluoroacetamide), as used in this thesis, and t-BDMS (tert-butyldimethylsilylation), as it reduces fragmentation and often showing the analyte's molecular ion peak [67]. Care must be taken, when annotating derivatised peaks as chemical side reactions, such as the conversion from arginine to ornithine, are possible [220].

A2.1.3.3. Retention time indices and retention time locking

In GC, the retention time is often expressed as a relative measure, a retention index, instead of absolute time. The retention index, first described by Ervin Kovats [336], is independent of the used GC system and therefore allows easy method transfers and database matching [27]. A retention index sets a compound's retention time in relation to the two alkanes in-between it elutes, when injected together with a series of gradually in chain length and therefore retention time increasing n-alkanes. According to equation A2 the Kovats retention index I depends on the retention time t_r and carbon chain length of the smaller (n) and larger (N) n-alkane.

$$I_{unknown} = 100 \times \left[n + (N - n) \frac{t_{r(unknown)} - t_{r(n)}}{t_{r(N)} - t_{r(n)}} \right] \quad (\text{A2})$$

However, as shown by Stein *et al.* [541], a retention index, similar to the Kovats retention index can also be calculated, based on the analyte's molecular properties, such as its functional groups, avoiding the use of n-alkanes as internal standards.

Another technique, recently developed and used in this thesis, to compensate for retention time deviations between different runs, batches or even systems is outlined by Blumberg *et al.* [54]. Retention time locking (RTL). Conveniently, in RTL only one internal standard is used for the global alignment of retention times and calculation of retention indices. This is done by calibrating the system's head pressure against the retention time of the internal standard. This enables a constant flow-rate, even though the gas expands during the temperature gradient [154].

A2.2. Analyte detection

Many mechanisms to detect analytes (after chromatographic separation) exist. One distinguishes between classical methods, such as titration or flame tests, and instrumental methods. Instrumental methods divide into spectroscopy, mass spectrometry, electrochemical methods and thermal methods. As the major focus of this PhD is on NMR spectroscopy and mass spectrometry only these two will be described in the following sections.

A2.2.1. Nuclear magnetic resonance (NMR) spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy relies on an atomic nucleus absorbing electromagnetic energy through excitation and re-emitting it at a specific, diagnostic frequency.

A2.2.1.1. The quantum physics behind NMR

In order for a nucleus to do so, it requires it to have a magnetic and angular momentum, which gives rise to its spin. A spin only exists if a an isotope has a spin number other than zero, *i.e.* an odd and/or even number of neutrons and protons. Some of the most frequent studied NMR active nuclei are for example ^1H , ^{14}N , ^{15}N , ^{17}O , ^{19}F or ^{31}P . However, this section will only deal with nuclei with a spin number of $I = 1/2$; ^1H and ^{13}C in particular.

According to quantum mechanics, a nucleus with a spin number of $I = 1/2$ can only have $2I + 1 = 2$ orientations. Once a nucleus is put into an external magnetic field B_0 , it either aligns with the field or against it, giving rise to a lower and higher energy level, respectively. This behaviour is displayed in figure A2.4. ΔE is the energy difference between those two spin states. According to the fundamental equation of spectroscopy (A3), a photon, matching this exact energy difference can excite the system by leveraging it from the lower to the higher state.

$$\Delta E = h\nu \tag{A3}$$

With h being Planck's constant ($h = 6.626 \times 10^{-34} \text{Js}$) and ν the nucleus resonance frequency. This frequency, also called Larmor frequency, is the frequency with which the individual nuclei rotate around their axis. This rotating motion is described by their spin.

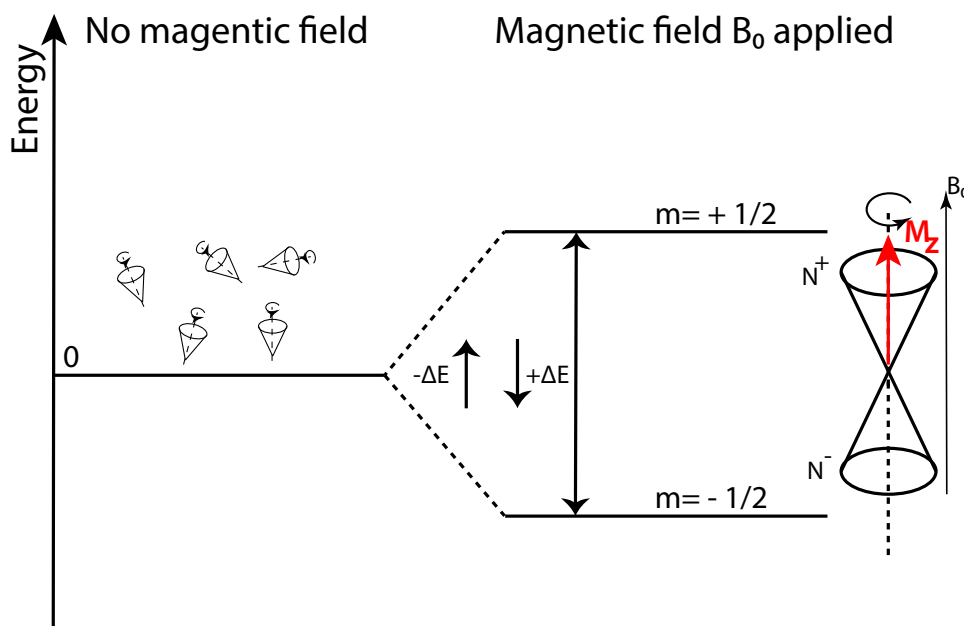


Figure A2.4.: **Spin physics of NMR spectroscopy :**

Upon applying a magnetic field B_0 the ^1H nuclei's spins align with the field, and precess around the B_0 axis, rather than random angles. Furthermore they divide into a lower (N^+) and higher (N^-), energy state, which can be reached via excitation with a photon of ΔE energy. M_z is the resulting net magnetisation vector between upper and lower spin state, which itself precesses with the resonance frequency parallel to B_0 .

In the case of the external magnetic field B_0 , the rotation axis is aligned with B_0 . B_0 does have a linear relationship with the Larmor frequency, expressed in equation A4.

$$v = \frac{1}{2\pi} \gamma B_0 \quad (\text{A4})$$

Therefore the Larmor frequency can simply be expressed as:

$$\Delta E = \frac{h\gamma B_0}{2\pi} \quad (\text{A5})$$

γ is the gyromagnetic ratio. It is a specific constant for each nucleus, such as 42.58 MHz/T and 10.71 MHz/T for ^1H and ^{13}C , respectively.

There are slightly more spins in the lower energy level (N^+), than the higher energy (N^-) level, giving the nuclei the net magnetisation vector M_z . The exact ratio is defined by

a Boltzmann statistic, according to equation A6, with k being the Boltzmann constant ($k = 1.3805 \times 10^{-23} \text{ J/K}$) and T the energy in Kelvin.

$$\frac{N^-}{N^+} = e^{-\Delta E/kT} \quad (\text{A6})$$

The ratio of the number of nuclei between the higher and lower excitation state directly defines the LOD of NMR spectroscopy. As both, the magnetic field strength and the gyromagnetic ratio are directly proportional to the excitation and therefore emitted energy ΔE , they both impact the LOD. Thus, some nuclei, despite equal isotopic abundance, are more sensitive than others. Also, stronger magnets will lead to higher sensitivity and improved resolution.

A2.2.1.2. The vector model of NMR spectroscopy

During every NMR experiment, the nuclei are excited into the higher state and the energy emitted upon relaxation is recorded. To excite the nucleus a 90° pulse is used. This is a high frequency electromagnetic pulse perpendicular to the net magnetisation vector M_Z , which drags it down from the Z into the X-Y plane (figure A2.5a). The net magnetisation is then allowed to rotate inside the X-Y plane at approximately the Larmor frequency. For convenience, another coordinate system, the rotating frame of reference system is used. The new coordinate system (X'-Y'), rotates inside the X-Y plane at the Larmor frequency, as shown in figure A2.5c. Any magnetisation vector that travels faster than the resonance frequency, travels clockwise about the Z-axis. Magnetisation vectors slower than the Larmor frequency move counter-clockwise (figure A2.5d).

When rotating inside the X-Y plane, the net magnetisation vectors induce a current into the detector coil, which gives rise to a signal, displayed in figure A2.5e. This signal, following a sinc-function, *i.e.* a (co)sine wave exponentially decaying to zero, is called free induction decay (FID). It is due to the relaxation of the excited nuclei. As the net magnetisation vector's magnitude rebuilds along its Z-axis, the current induced into the detector coil ceases. The magnetisation vector's relaxation from zero magnitude in the Z-direction to its original (unexcited) M_z value is called the spin-lattice relaxation (T1). The reduction of the vector's XY magnitude is referred to as spin-spin relaxation (T2) and is always less or equal the time required for the spin-lattice relaxation. The magnetisation decay is greatly

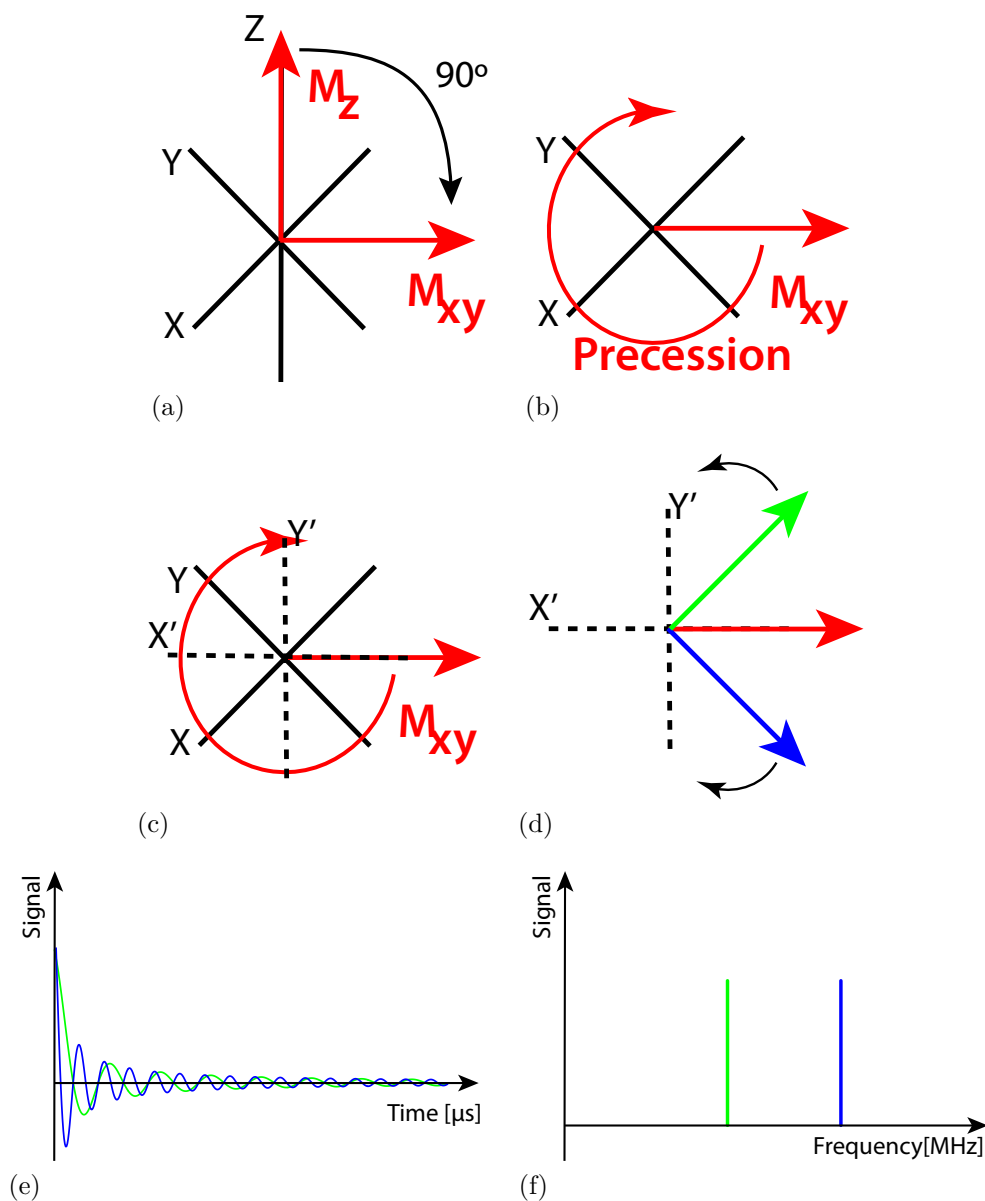


Figure A2.5.: **Vector model of NMR spectroscopy:**

The M_z net magnetisation vector is tilted into the X-Y plane by a 90° pulse (figure A2.5a), inside which it keeps processing at the resonance frequency (figure A2.5b). A X'-Y', rotating frame coordinate system is defined, rotating at the resonance frequency (figure A2.5c). In this new coordinate system, spin from nuclei precessing faster than the resonance frequency travel clockwise (blue), the ones slower, counter-clockwise (green) (figure A2.5d). The precession is detected before relaxation into the unexcited state. The decaying signal is called free induction decay (FID) (figure A2.5e). After Fourier transform, a frequency spectrum is yielded, showing the resonance frequency of the two different nuclei populations. [57]

governed by two factors; molecular interactions and inhomogeneities in B_0 .

Figure A2.5e also shows that the two different net magnetisation vectors having a different frequency during their decay. This is due to their slight deviation from the resonance frequency. When the FID is converted from the time to the frequency domain, using a Fourier transform, a frequency spectrum, displayed in figure A2.5f is yielded. [103, 360]

A2.2.1.3. The chemical shift

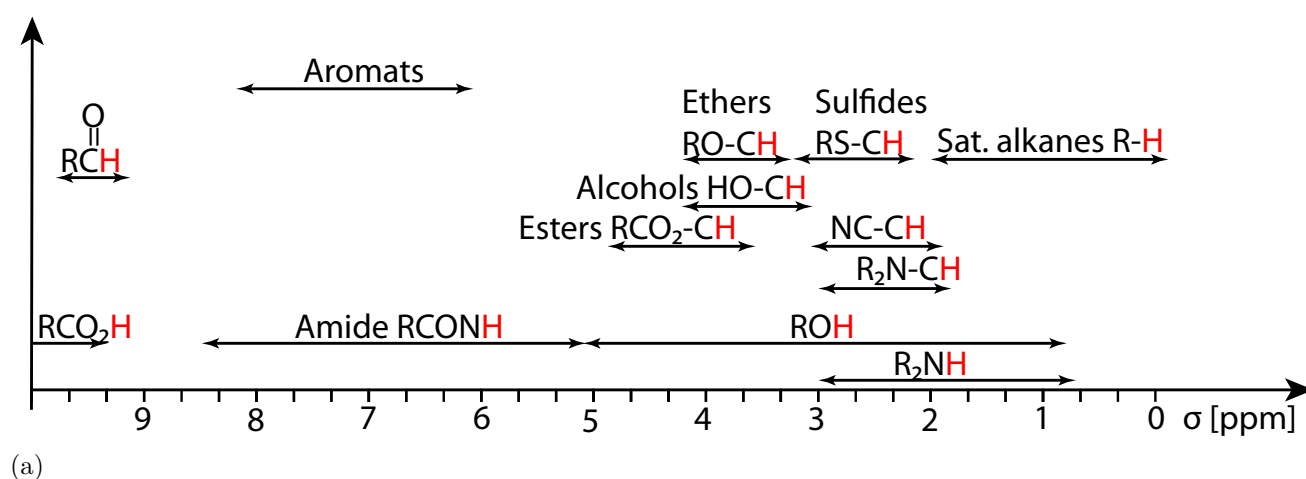


Figure A2.6.: ^1H NMR chemical shift range of different functional groups:

The chemical proton shifts from various functional groups are given on the ppm scale.

These small deviations from the resonance frequency are referred to as chemical shift. They provide diagnostic information on the chemical environment of the nucleus. Each nucleus does not directly experience the native B_0 field, as the atom's circulating electrons generate a small magnetic field themselves, which shields the nucleus from B_0 . According to equation A5 a small change in B_0 does directly affect ΔE and therefore the precision frequency, leading to the chemical shift. The electrons and therefore the chemical shift is influenced by the type of bond and neighbouring chemical groups of different electronegativity. Most chemical functional groups have specific shift ranges [442]. Figure A2.6 shows some common examples. As the spectrometer's resonance frequency depends on its magnetic field strength, using a Herz scale would make it difficult to compare between

spectrometers. To this end, a relative scale is used. Each resonance is scaled relative to a chemical standard, according to equation A7, to yield the chemical shift σ , expressed as ppm. As chemical standards, tetramethylsilane (TMS) and its organic acid derivatives such as trimethylsilyl-2,2,3,3-tetradeuteriopropionic acid (TSP) are used for organic and aqueous solvents, respectively. [252]

$$\sigma = (v_{\text{Sample}} - v_{\text{Reference}}) \times \frac{10^6}{v_{\text{Reference}}} \quad (\text{A7})$$

A2.2.1.4. Spin-Spin coupling

Another NMR specific phenomenon is spin-spin coupling. [360] Not only do close functional groups impact the magnetic micro-environment of a proton, but also its closest proton neighbours, *i.e.* those connected over 2 to 3 bonds. The interaction of two neighbouring protons, that are in different chemical environments, *i.e.* have different chemical shift, leads to signal splitting. Signal splitting leads to multiplets, such as doublets, triplets, quartets, etc. In a simple spin system, *i.e.* when a proton only couples with one chemically different group, it has a multiplicity of $n + 1$ with n being the coupling partner's proton count. The intensity distribution of the multiplet's peaks follows Pascal's triangle, according to the number of possible quantum states. Figure A2.7 illustrates this principle by using ethanol ($\text{C}_2\text{H}_5\text{OH}$) as an example.

If a proton can see two or more protons of different chemical environments or stereochemistry, which also introduces non-equivalence, higher order splitting patterns result. For example the ethyl group in the middle of 3-Amino-1-propanol ($\text{NH}_2 - \text{CH}_2 - \text{CH}_2 - \text{CH}_2 - \text{OH}$) has two chemically different methyl groups to either side. It will therefore see two triplets, which results in a triplet of a triplet multiplicity, with a complex intensity pattern.

Due to the ppm scale, a multiplet's chemical shift, which is located at the symmetry axis of the multiplet is independent of an NMR's field strength. The coupling constant J , however, is not. J is the difference between the individual maxima of the multiplet. It is measured in Hz and is a measure of the magnetic interaction between the coupling nuclei. Figure A2.8 explains the coupling constant and multiplicity (doublet, triplet, quartet) from a vector model perspective. In the previous ethanol example, both, the triplet and the quartet have

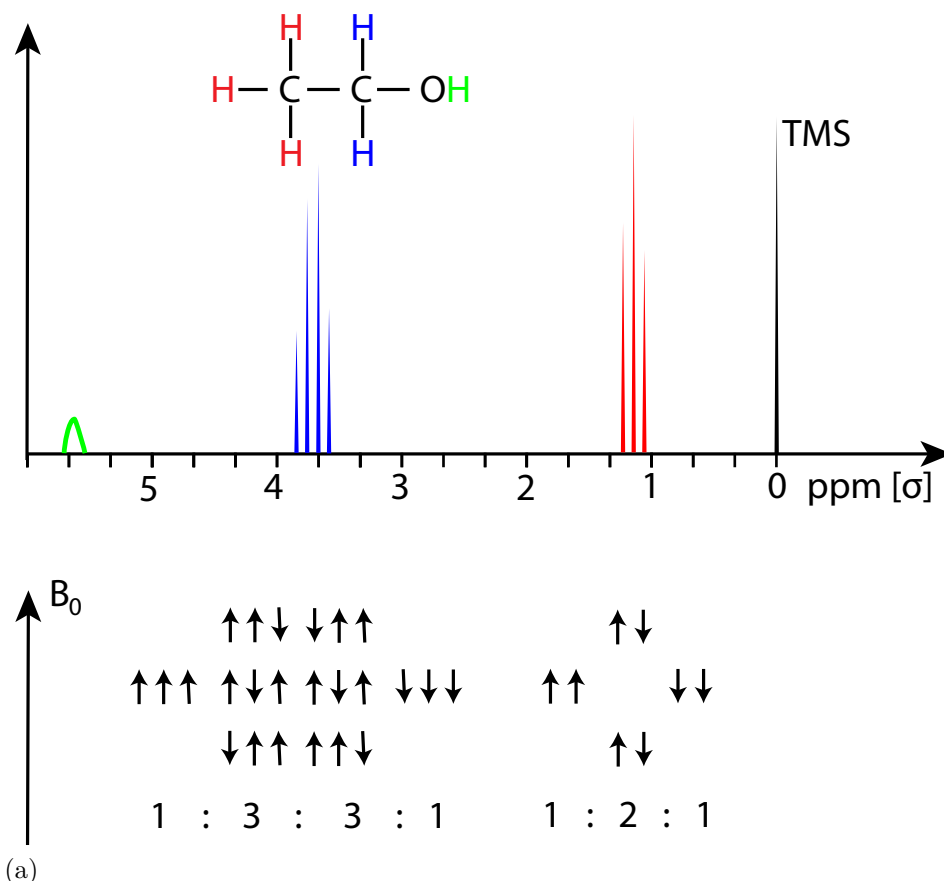


Figure A2.7.: NMR spectrum of ethanol and the corresponding spin populations:

Ethanol's methyl group (CH_3) can "see" three different spin orientations when coupling with its neighbouring ethyl group (CH_2). As two of the spin states are effectively the same (up/down), those have twice the intensity as the other two, resulting in a 1:2:1 signal intensity. The methyl group, however, can see its ethyl partner in four different spin states, with the intensity ratio of 1:3:3:1 according to Pascal's triangle. The CH_3 protons of the spiked in chemical shift reference TMS, however, have no coupling partner and therefore appear as singlet. Even though the alcohol's acidic proton was illustrated in this spectrum, it usually is not visible, if the used solvent, *e.g.* D_2O allows for proton/deuterium exchange. Furthermore, the methyl group and acidic proton do not produce a multiplicity as there is no coupling across heteroatoms such as O, N or S.

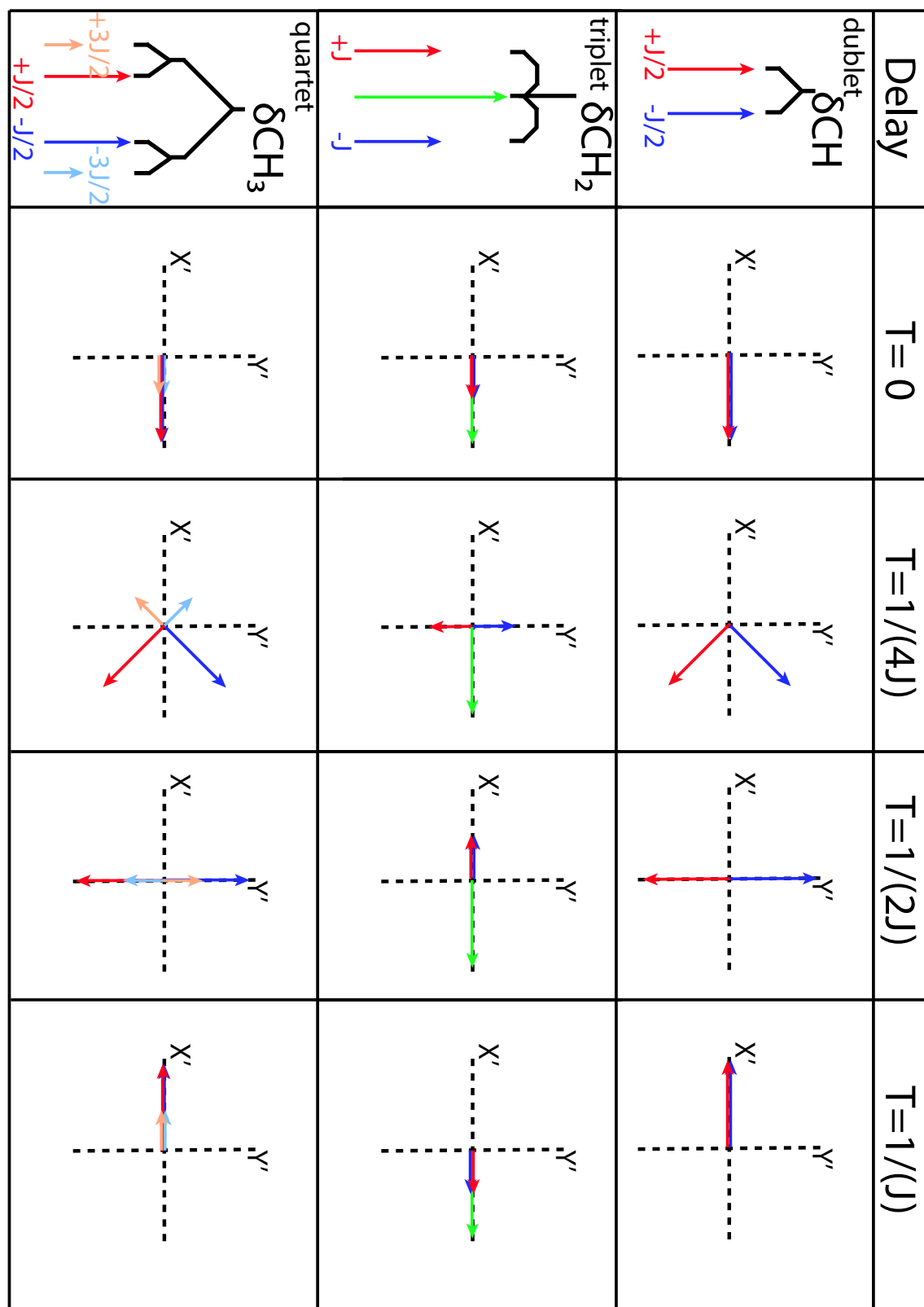


Figure A2.8.: Chemical shift and coupling in the NMR vector model

a coupling constant of $J = 7$ Hz. The coupling constant can also be used as diagnostic criterion, as just like the chemical shift, it is specific for different chemical groups. It even allows statements about the conformational interaction between two protons, such as the dihedral angle.

Chapter III. builds on a 2D NMR sequence in which carbon-carbon couplings are "removed" from the spectral area, by collapsing their coupling constants. This can be achieved by a 180° pulse, as shown in figure A2.9.

The 180° is also a vital part of other pulse sequences, as it allows to refocus the spins. This is important as due to inhomogenous magnetic fields, the spin isochromats may dephase and lead to undesired line broadening. A 180° refocusing pulse can reverse this.

A2.2.1.5. NMR for quantification

NMR is a quantitative detection technique. The integral of a multiplet is proportional to the number of its underlying protons in a given molecule and sample. The multiplet integral ratio of the methyl:ethyl in ethanol should therefore be 3:2. If an internal standard of known concentration and proton number is used, its integral can be used to calculate the absolute concentration of any other analyte in the solution. However, for precise, fully quantitative work, several parameters require careful consideration [506, 212].

A2.2.1.6. NMR instrumentation

To perform NMR experiments, an NMR instrument, comprising of the superconducting magnet, the probe and various electronic systems, is needed. The principal assembly of a magnet and detector probe is illustrated in figure A2.10.

To achieve field strengths of several hundred MHz, extremely high magnetic field strengths are needed. For example 14.1 Tesla are required for 600 MHz ^1H observation frequency. Such magnetic fields can only be supplied by superconducting solenoid magnets. For the super conduction effect cryogenic temperatures are required. Therefore, the magnet is chilled by liquid helium (4.2 °K), which itself is cooled by liquid nitrogen (77 °K), separated from the atmosphere by a vacuum dewar. Both cryogenics require regular top-ups. A sample tube is introduced into the magnet via a sample port and kept spinning around its axis by a temperature controlled stream of air, inside the probe. Spinning the sam-

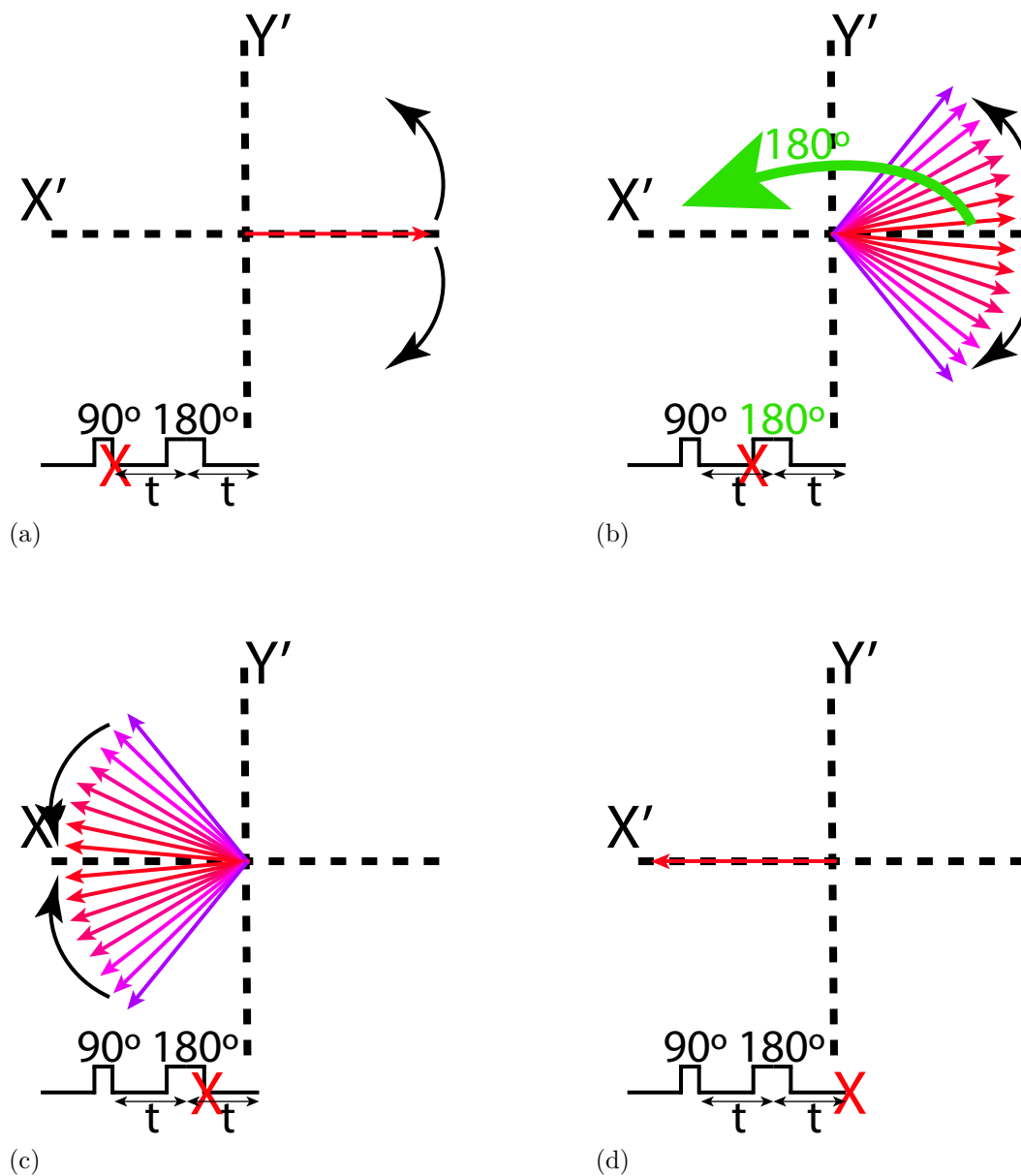


Figure A2.9.: **The 180° refocusing pulse:**

A 180° pulse is an important building block in the design of pulse sequences, as it allows to refocus the couplings. After a 90° pulse excites the spins into the $X'Y'$ plane (figure A2.9a), they start to precess (and dephase). After a defined time t , a 180° pulse along Y' is applied (figure A2.9b). They individual spins are mirrored with respect to the Y' axis, but proceed precessing with their initial direction and speed (figure A2.9c). After the time t , all spins will be refocussed 180° from their starting position (figure A2.9d).

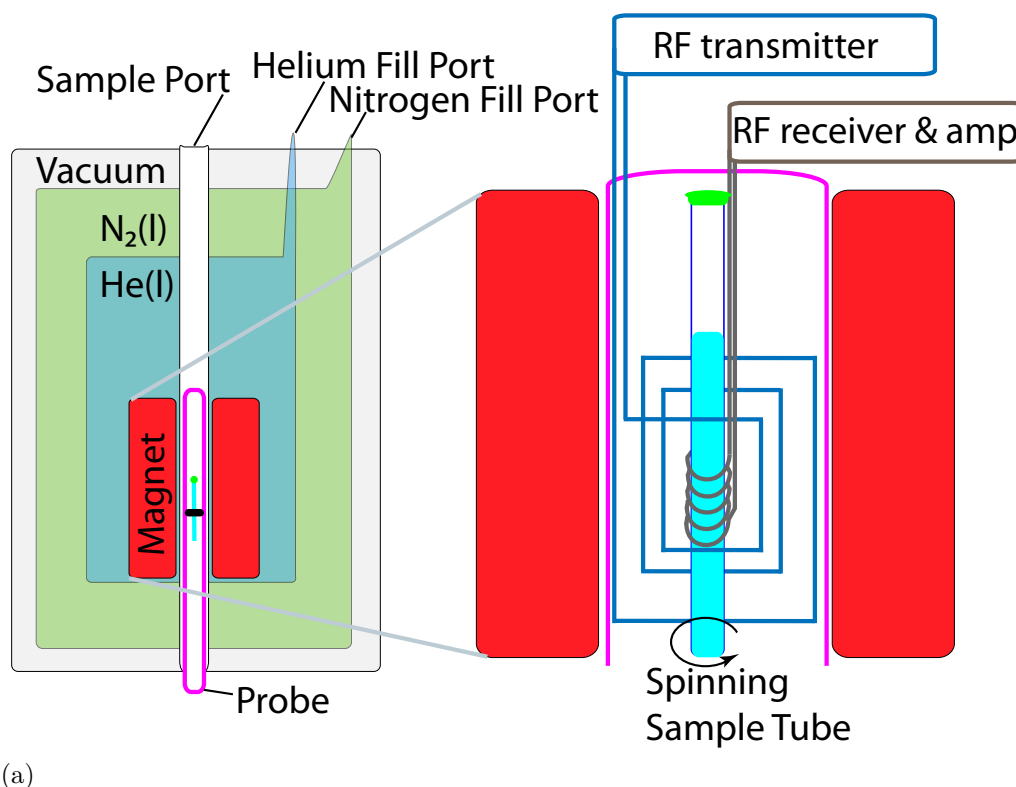


Figure A2.10.: **Assembly of a NMR magnet and probe:**

Assembly of the liquid helium and liquid nitrogen chilled superconducting magnet on the left. The probe is magnified on the right. For explanations refer to the text.

ple averages out minor field inhomogeneities and allows for a narrower line width. The specifics of the probe determine the bore of the NMR tubes (5 mm in this thesis) and the observable nuclei. Inside the probe, there is an radio-frequency (RF) coil to irradiate the sample with the pulse sequence and a RF detector coil, to sample the FID. Both are controlled by electronics. The detector coil feeds the signal through various filters into an amplifier which passes the signal on to an analogue-to-digital converter (ADC), that is connected to a computer.

NMR spectroscopy has a relatively high LOD. One method to improve the signal to noise ratio (SNR) and therefore LOD of acquisitions is to maintain the receiver RF coils and preamplifier at very low temperatures. This reduces the noise contribution from the random thermal motion of electrons inside the conductors (Johnson-Nyquist noise) and therefore improves the SNR up to five fold. Some experiments in this thesis were performed using

low concentration samples and were therefore acquired with such a "cryo-probe". [549]

Metabolomics sample sets often consist of several tens of samples. Therefore automatic sample handlers are frequently used. In order to reduce effort and ensure consistency, an automatic sample manager was used for certain experiments in this thesis. The sample manager changes samples and automatically performs the shimming and receiver gain adjustment, as outlined in the next section.

A2.2.1.7. NMR sample set acquisitions

The standard procedure of an NMR sample (set) acquisition is outlined in this section. Figure A2.11 outlines all major steps.

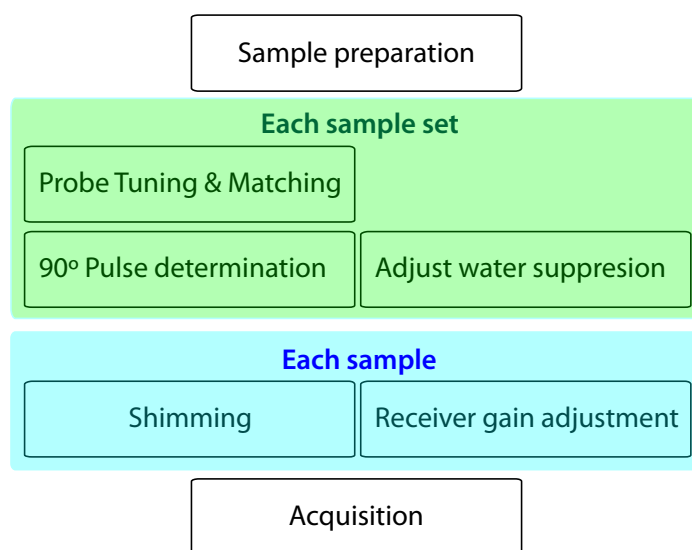


Figure A2.11.: **Standard procedure of a NMR sample acquisition :**

Flow chart of a standard procedure for a NMR acquisition [36] of a sample set, used in this thesis. For details please refer to the text.

An NMR experiments starts with sample preparation. Dried samples have to be resuspended in appropriate solvents. For two reasons, those solvents are usually deuterated. First, the NMR uses deuterium as a "lock", in order to correct for frequency drifts and temporary magnetic field inhomogeneities. Secondly, the solvent's protons would be NMR visible themselves. As compared to the analyte, the solvent is very highly concentrated, the

dynamic range of the ADC would not be able to grasp a reasonable signal of the analyte. Therefore a 100 % D_2O pH 7 phosphate buffer with 0.5 mM TSP as shift reference will be used for all NMR experiments in this thesis.

Before starting a series of acquisitions (particularly after using a different solvent), the magnets resonance frequency requires to be tuned and matched. Matching and tuning change the impedance of the probe and adjust the resonance frequency of the RF coil, respectively. Both parameters depend on the polarity and dielectric constant of the used solvent composition and are essential to minimise the power needed for a 90° pulse. After the sample has been introduced and allowed to equilibrate to the set temperature, the 90° pulse for the given matrix can be "found". This involves finding a RF pulse length (at a set pulse power), that flips the majority of the nuclei by 90° , (into the XY plane) to reach the maximum signal amplitude. In practice, the 360° pulse, *i.e.* the pulse at which the net magnetisation vector is back at the starting position, orthogonal to the detector coil, with therefore no visible signal, is determined and divided by 4. Finally, the water suppression must be adjusted. Suppressing the a residual water peak is particularly important when analysing low concentration samples, as is the case within the aims of this thesis. The used pre saturation water suppression sequence used in this thesis irradiates a RF signal prior to the acquisition to saturate and silence all H_2O and HDO protons [529]. In order to suppress the water peak effectively and not affect neighbouring regions, the RF frequency has to be adjusted to the apex of the residual solvent peak.

Once all the parameters have been found, the desired pulse sequence can be set up and specific parameters adjusted as desired. An important parameter is the number of scans. An NMR experiment, consisting of the pulse train and FID acquisition, is usually conducted several times and the individual spectra averaged to improve the signal to noise ratio (SNR). According to equation A8, the SNR is proportional to the square root of the number of individual acquisitions (scans). This means doubling the number of scans will double acquisition time, but only improve the SNR by 1.4 fold. Therefore a reasonable tradeoff is required. For pulse sequence specific reasons (phase recycling) it often is recommended to set the number of scans to the multiple of a certain factor, such as 4.

$$\text{SNR} \propto \sqrt{n_{\text{Scans}}} \quad (\text{A8})$$

The next two steps prior to acquisition, the shimming and receiver gain adjustment, need to be performed with each individual sample. During the shimming the field strength of small magnets inside the superconductor's B_0 field are (manually) adjusted to even out field inhomogeneities. Low field homogeneity leads to asymmetric or too wide peak line widths, reducing resolution. As each sample disturbs the B_0 homogeneity it is essential to shim and check the line width prior to each acquisition. During the receiver gain adjustment, the optimal gain for the amplifier, to use the full dynamic range of the ADC, is determined. Too low receiver gains compromise the SNR, too high receiver gains lead to signal distortions and artefacts.

After all these steps, the acquisition can be started. After the acquisition, the next sample is inserted, shimmed, receiver gain adjusted and acquired. Details on the subsequent processing of NMR spectra is discussed in section A3.1.1. [36, 57]

A2.2.1.8. Two dimensional NMR spectroscopy

Although most NMR based metabolite profiling approaches use one dimensional ^1H pulse sequences, a few examples of two dimensional NMR based profiling have been published [376, 15]. However, two dimensional (2D) NMR is widely used for assignment purposes. The sequences used for metabolite annotation in this thesis will be described briefly. Two major types of experiments were used; ^1H - ^1H homonuclear and ^1H - ^{13}C heteronuclear experiments.

^1H , ^1H - Correlation Spectroscopy (COSY) was the predominant used homonuclear pulse sequence. COSY maps an ^1H on its X and Y axis. The cross peaks correlate the neighbouring protons. In the case of ethanol, the triplet and quartet of the ethyl and methyl would be connected via a cross peak. This is very useful to assign spin systems that could not be correlated just from a one dimensional spectrum. A COSY works by transferring magnetisation from one spin to its coupling partner (coherence transfer), yielding a spectrum which is symmetric with regards to its diagonal.

Heteronuclear Single Quantum Coherence (HSQC) spectroscopy is the most frequently used heteronuclear pulse sequence. It allows to correlate ^1H signals to the ones of ^{13}C or other heteroatoms, such as ^{15}N . Therefore its cross peaks indicate the heteroatom to

which a proton is attached. This is not only helpful for de-novo elucidation, but also for assignment as it introduces a second, (independent) variable, the carbon ppm shift. As the carbon shift range ranges from 0 up to over 200 ppm it can be deciding for ambiguous ^1H shifts. Also, as it spans from 0 - 10 $\times$ 0 - 200, spectral overlaps occurring in the ^1H domain can often be resolved. However, it has a major draw back. Its long acquisition times and low inherent sensitivity, due to the lower gyromagnetic ratio of carbon and the only 1 % natural abundance of ^{13}C . Chapter III. aims at addressing this issue by using performing acquisitions with ^{13}C enriched samples.

Heteronuclear multiple-bond correlation spectroscopy (HMBC) extends HSQC to long range couplings, in the order of 2 - 4 bonds. It is mainly useful for structure elucidation as it allows to connect several spin systems.

There are several other two dimensional established pulse sequences, such as total correlation spectroscopy (TOCSY) or J-resolved spectroscopy (JRES). TOCSY extends COSY up to 5 bonds, whereas JRES directly measures coupling constants and correlates these to the multiplets from the one dimensional NMR spectrum. However, neither of these techniques were used in this thesis.

A2.2.2. Mass spectrometry

Mass spectrometry detected the charge to mass ratio (m/z) of ions. Figure A2.12 shows a mass spectrum of ethanol.

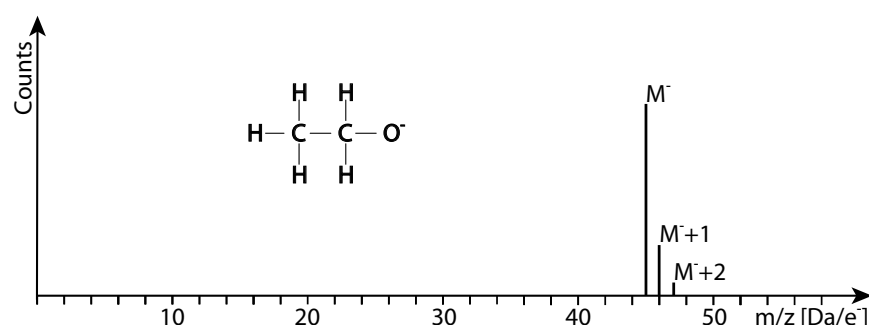


Figure A2.12.: **Mass spectrum of ethanol:**

Mass spectrum of ethanol. The negatively ionised M^- molecule has a m/z of 45, followed by its $M + 1$ and $M + 2$ isotope peaks.

All mass spectrometers consist of three basic units, outlined in figure A2.13.

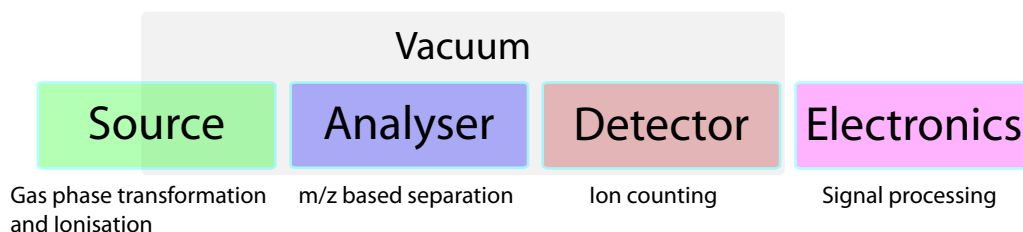


Figure A2.13.: **Basic assembly of a mass spectrometer:**

The basic units of a mass spectrometer will be described in the text.

First, inside the source, the sample is transferred into the gas phase and analytes are ionised. Secondly, ions are separated according to their mass to charge ratio (m/z) inside a mass analyser. Finally, the ions are counted by a detector unit. Although rarely mentioned, depending on the instrument type there is a fourth unit, consisting of sophisticated signal processing modules. Furthermore, as starting at the source, most of these units require a (ultrahigh) vacuum to avoid collisions of the ions with gas atoms and remove non-ionised matter. The vacuum, down to 10^{-10} Torr is often built up over several stages using rotary and (turbo)molecular pumps.

A2.2.2.1. Types of ionisation sources

A plethora of different ionisation techniques have been invented and are available for mass spectrometers. The choice of source depends on the type of analytes and matrix to be analysed [504, 262]. Even though not clear cut, ionisation techniques may be classified into several categories. Electron impact and chemical ionisation, gas discharge ion sources, desorption ionisation, spray ionisation and ambient ionisation.

EI (Electron Impact ionisation) [134] and CI (Chemical Ionisation) [430] are predominantly used for GC-MS analysis and will be discussed in section A2.3.3. ICP (Inductively Coupled Plasma) [253] may be the most widely used gas discharge ion sources, particularly as ICP-MS is the method of choice for trace metal analysis. Desorption ionisation techniques include LMIS (Liquid Metal Ion Source) [554] mostly used by SIMS (Secondary Ion Mass Spectrometry) and FAB (Fast Atom Bombardment), pioneered by Barber and Morris [420].

Remarkably, due to its soft ionisation characteristics, FAB was the first technique to allow the analysis of whole biomolecules such as insulin [32]. The spray ionisation technique used for the LC-MS experiments in this thesis was ESI (Electrospray Ionisation), pioneered by Fenn *et al.*[170]. It will be described in detail in the LC-MS section (A2.3.2). The other notable spray ionisation source is APCI (Atmospheric Pressure Chemical Ionisation). Among the ambient ionisation techniques MADLI (Matrix-Assisted Laser Desorption Ionisation), simultaneously developed by Tanaka *et al.*[562] and Hillenkamp *et al.*[300] is very prominent in the field of protein and (bio)polymer analysis. More frequent developments include DESI (Desorption Electrospray Ionization) and DART (Direct Analysis in Real Time). Recently, ambient ionisation techniques have gained particular popularity, as those enable mass spectrometry imaging (MSI) approaches. MSI rasters the surface of a sample and produces an image; however, instead of colour, like in microscopy, each pixel is associated with a mass spectrum and therefore chemical information [545].

Most of these techniques are able to produce positively (cations) and negatively (anions) charged ions, respectively.

A2.2.2.2. Types of mass analysers

Once all ions have been transferred to the gas phase, they need to be separated based on their mass to charge ratio, prior to detection. As for ionisation sources, several different types of mass analysers, based on different separation principles, exist. Each with their own unique properties and therefore areas of application. This section discusses the basic properties of mass analysers, before listing them. The ones used in this thesis, the time-of-flight (ToF) and quadrupole (Q) analyser will be discussed in more detail.

Mass accuracy is a mass analyser's ability to deliver accurate and precise measurements.

Most mass analysers default into two categories. Low and high mass resolution. Low mass resolution instruments can only discriminate between integer differences in molecular mass. High mass resolution instruments can differentiate between significantly closer related masses. Their accuracy is measured as the relative measurement error, in relation to the measured mass, expressed as parts per million (ppm). Equation A9 defines the ppm error.

$$Error_{ppm} = \frac{(mass_{measured} - mass_{calculated})}{mass_{calculated}} \times 10^6 \quad [ppm] \quad (A9)$$

High mass accuracy usually requires the continuous infusion of a reference standard of known m/z ratio, to correct for subtle drifts throughout the run. For this purpose, a Leucine-Enkephaline solution will be used as lock spray in this thesis.

Resolution or resolving power is the mass analyser's ability to distinguish two closely related m/z peaks. According to equation A10 it is defined as the fraction between a measured m/z value and Δm , which usually is the full width half maximum (FWHM), *i.e.* the peak's width at half its height.

$$R = \frac{m}{\Delta m} \quad (A10)$$

Mass Range determines the m/z range within peaks can be measured. This value is more likely to have an upper, rather than a lower limit.

Dynamic Range is the order of magnitude in difference between the largest and smallest quantity that can simultaneously be determined.

Scan rate is the rate at which mass spectra are acquired. This value is of particular importance if the changing sample composition throughout the acquisition is of interest; as in the case of an online-coupled separation step (LC-MS; GC-MS).

Table A2.1 outlines the most common mass analysers and some of their basic properties, that will be discussed in the following text.

Analyser	Resolution [$m/\Delta m$ @ FWHM]	Mass accuracy [ppm]	Scan rate [scans/s]
Sector	3 000 - 80 000	≤ 5	0.1 - 0.5
Quadrupole	$\leq 3\,000$	low	0.5 - 4
Ion trap	$\leq 4\,000$	low	5 - 10
Time-of-flight	8 000 - 40 000	1 - 10	20
Orbitrap	$\leq 100\,000$	≤ 1	0.5 - 2
FT-ICR	$\leq 1\,000\,000$	≤ 1	0.3 - 1

Table A2.1.: Basic properties of different mass analysers. [619, 130] Exact values depend on the instrumental setup and vary between instrument manufacturers.

(Electro/Magnetic) sector instruments were the first type of mass spectrometers [395, 284] and are well suited to demonstrate the general separation principle of a mass spectrometer. When moving inside a electric or magnetic field, charged particles experience a Lorentz force F . According to equation A11 this force depends on the ion's charge z , the electric field E and the cross product of the ion's velocity v and the magnetic field B .

$$F = z(E + v \times B) \quad (\text{A11})$$

According to Newton's second law (equation A12) the force F is required to accelerate a particle of the mass m , with an acceleration of a .

$$F = ma \quad (\text{A12})$$

Combining both equation, one yields equation A13

$$m = \frac{z(E + v \times B)}{a} \quad (\text{A13})$$

Considering the elementary charge of an electron ($z = e = 1.602 \times 10^{-19}C$), a differential equation A14 is yielded, that is fundamental to any mass analyser. It sets an ion's m/z in relation to space and time, the two domains all mass analysers use to perform separations.

$$m/z = \frac{(E + v \times B)}{a} \quad (\text{A14})$$

Sector instruments have a fixed magnetic and/or electric field of known strength. If only one, a magnetic or an electric field is used, the above equation can be simplified by omitting the corresponding term; E for magnetic sector and $v \times B$ for electric sector instruments). Following from the equation and the right-hand rule, the ion is deflected on a path with a radius proportional to its m/z . A detector can scan the path and determine a particle's m/z . Albeit doubly focused instruments (an electrostatic analyser is added) achieve good accuracy and resolution, their slow scanning speeds, large footprint and high cost has widely led them to be replaced with other mass analysers.

Quadrupole mass analysers [468] consist of four circular, parallel arranged rods, as displayed in figure A2.14. A RF alternating current (AC) voltage, with a direct current (DC) offset is applied to the four rods. The same ACDC voltage is always applied to the two opposing rods. The AC proportional for the non-oppositely situated rods is 180° phase shifted and therefore having an opposing charge. Ions in-between the four rods will expe-

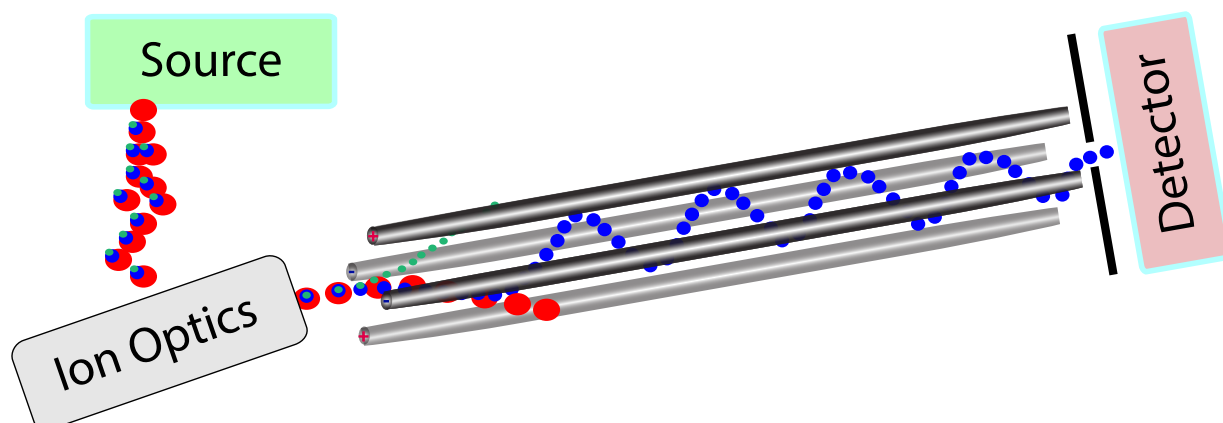


Figure A2.14.: **Quadrupole mass analyser:**

Ions of different m/z ratios from the source, focused by ion optics, are accelerated into four parallel rods. By applying a complex voltage pattern and therefore generating a quadrupolar electric field in-between the rods, ions are kept in this field, moving towards the detector in a spiral motion. By controlling the generated field, ions of particular a m/z ratio can be filtered out prior to reaching the detector.

rience a quadrupolar electromagnetic field. Any ion accelerated to travel parallel to the rods, in the gap between the rods will travel along those in a spiral motion. By varying the AC to DC ratio, the acceleration and speed of the ions along the spiral path can be controlled. According to the differential equation derived above, this allows masses with certain m/z to be selected. Indeed, as illustrated in figure A2.14, the quad operates like a m/z filter, only allowing a selected range of ions to reach the detector stage. Other ions, with unstable trajectory will collide with the rods and are therefore be removed [130, 10]. During regular operation the quadrupole "scans" along the m/z range during each scan to yield a mass spectrum. Therefore, if used as a "detector" for a large range of masses, it has a non-optimal LOD as the dwell time, *i.e.* the time spent to allow ions to pass the quad must be shared between all the m/z bins in every scan. Also the mass accuracy and resolution and scan rate are the lower end of mass analysers. However, if used as a filter or to monitor very few ions at a time, like in Q-ToF or triple-quad instrumentation, it can achieve a very low LOD/LOQ. Those hybrid instruments are discussed in section A2.2.2.4. As it is a very inexpensive and robust mass analyser, it is also the most used mass analyser for GC-MS.

Ion trap analysers [467], were invented by Wolfgang Paul, just like the quadrupole mass

analyser, which it shares most features with. An ion trap uses the quadrupole field to store ions of selected masses, by keeping them on a stable trajectory. Masses can be ejected selectively for detection or fragmentation. It is possible to collect the fragments and repeat the process of selected precursor fragmentation (MS^n). This technique has become essential in the mass spectrometric analysis of proteins.

Orbitrap mass analysers are a very new, but already important type of trap [387]. Ions are trapped by the balance of the attracting electrostatic field around a spindle-shaped electrode and their experienced centrifugal forces. They orbit this spindle in elliptical trajectories, with an angular frequency ω according to equation A15, with k as a force constant. The harmonic oscillation around the spindle shaped electrode induces a current into a detector, which can be converted into a m/z spectrum.

$$\omega = \sqrt{\frac{1}{k} \times (m/z)} \quad (\text{A15})$$

A particular advantage of this trap over a quadrupole trap is its FTICR like performance with regards to mass accuracy, resolution and sensitivity [260].

FTICR (Fourier Transform Ion Cyclotron Resonance) [109] mass analysers are considered the most accurate and highest resolution mass analysers. Ions are trapped in a Penning trap, *i.e.* inside a magnetic field created by a strong NMR-like superconducting magnet. The circularly rotating ions are excited by an orthogonal electric field (at their cyclotron resonance frequencies) to rotate in a larger cyclotron radius. After removing the electric field, the ions rotate at their cyclotron resonance frequency f , according to equation A16. Upon passing the detector electrodes, they produce a free induction decay (FID), which can be Fourier transformed into a mass spectrum. Like for the Orbitrap, mass resolution and accuracy is inversely proportional to scan rate, requiring a trade-off setting if coupled to an LC system [196].

$$f = \frac{Bz}{2\pi m} \quad (\text{A16})$$

Time-of-flight (ToF) [82, 633] mass analysers achieve a mass accuracy rivalling the one of the orbitrap and FTICR analysers. However, the scan rate is significantly better, and instrument costs lower, making it an attractive choice for metabolite profiling applications. Indeed, ToF based instruments will be used for detection in the LC-MS experiments of this

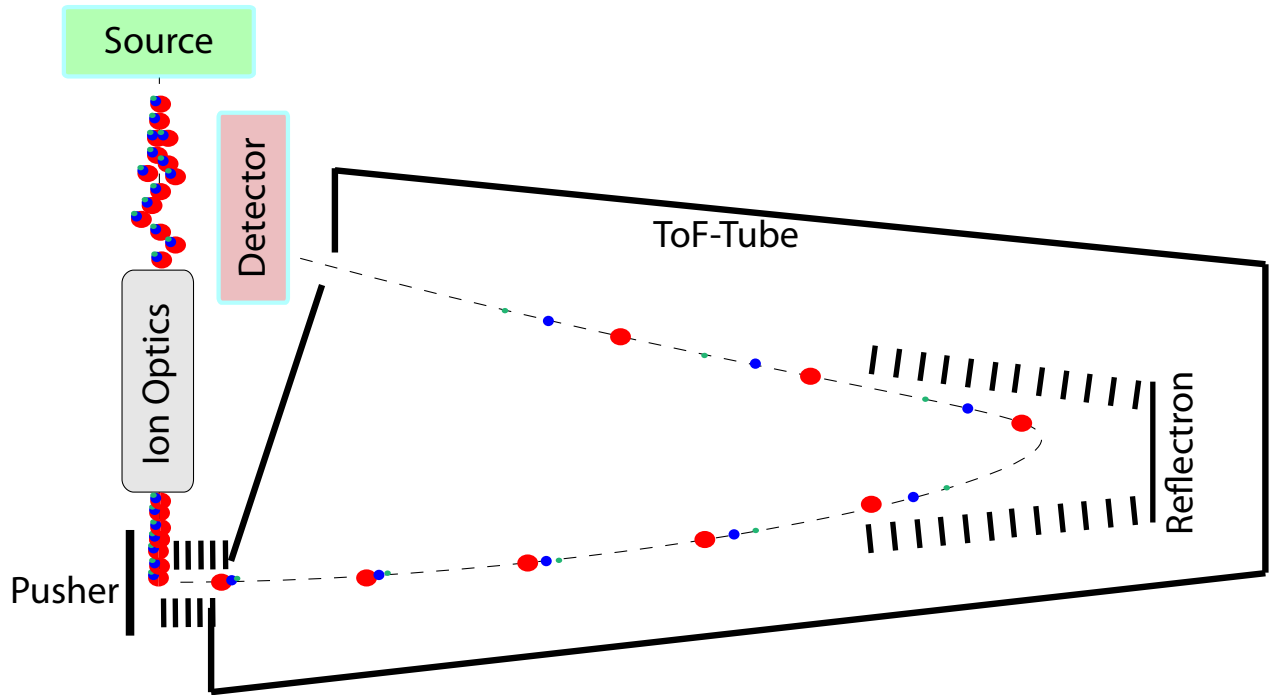


Figure A2.15.: **Time-of-flight (ToF) mass analyser:**

Ions of different m/z ratios from the source, focused by ion optics, are accelerated into the ToF tube by an ion-pusher. Throughout the flightpath, the lower speed of ions with a higher m/z ratio causes them to fall behind the ones with a smaller m/z ratio. The detector registers the time an ion needed to complete the flight path, which is proportional to its m/z . An ion-mirror like reflectron doubles the flight path.

thesis. Figure A2.15 shows the fundamental principle of a ToF analyser. An electric field with the (voltage) potential U , is generated by an ion-pusher, to accelerate a particle into the ToF-tube. According to equation A17, a particle of the charge z , experiences the potential energy E_p , when passing through the (voltage) potential U .

$$E_p = zU \quad (\text{A17})$$

The whole potential energy E_p acquired through the field, leads to an acceleration of the particle and will be completely converted into kinetic energy along the way (equation A18).

$$E_k = E_p \quad (\text{A18})$$

According to equation A19, the kinetic energy E_k depends on the mass m and speed v of

an object.

$$E_k = \frac{1}{2}mv^2 \quad (\text{A19})$$

Therefore, after the field, the ion will have a kinetic energy according to equation A20. As the provided energy is the same, ions with higher mass will have gained less speed.

$$zU = \frac{1}{2}mv^2 \quad (\text{A20})$$

Considering that after the acceleration through the potential the ion's speed remains constant and that according to equation A21 speed is the time t an object needs to travel the distance d , equation A20 can be substituted to yield equation A22.

$$v = \frac{d}{t} \quad (\text{A21})$$

$$t^2 = \frac{d^2}{2U} \sqrt{\frac{m}{z}} \quad (\text{A22})$$

Equation A22 establishes a relationship between the ion's flight path length d and its m/z ratio. After taking the square root and considering that according to equation A23, the constant potential U and flight path length d can be summarised in the constant k , one yields a final equation A24.

$$k = \frac{d}{\sqrt{2U}} \quad (\text{A23})$$

$$t = k \sqrt{\frac{m}{z}} \quad (\text{A24})$$

Equation A24 displays a simple relationship between the flight time t and the square root of the m/z ratio. In other words, by registering the time an ion took to travel along the ToF tube, its m/z ratio can be calculated. This is easily possible. At $t = 0$ the pusher (, displayed in figure A2.15) accelerates an ion package onto the flight path. The detector registers at which time how many ions arrive. This count versus time distribution can then be converted into a count versus m/z distribution - a mass spectrum.

The ToF analyser in figure A2.15) has a reflectron [9, 390], instead of having a linear flight path from pusher to detector. A reflectron acts as an ion mirror and is used to adjust for the kinetic energy dispersion of the ions. Furthermore, doubling the flight path improves

mass resolution as ions have more time to separate. However, it inversely affects sensitivity, as ions may get lost on the way. The illustrated reflectron is operated in "V-mode". Modern mass spectrometers can also operate in "W-mode", using a second reflectron to even further increase the path length. This leaves the operator the choice between resolution and sensitivity.

It is also possible to combine mass analysers by operating them in series. Two very common examples are the triple quad (QQQ) and the Q-ToF. The former is a linear combination of three quadrupole mass analysers, the second a combination of a quadrupole and a time-of-flight analyser. Both of them have major importance for molecule fragmentation studies, as explained in section A2.2.2.4.

A2.2.2.3. The Detector

After all ions have been separated according to their m/z , they require to be registered by a detector, before a mass spectrum can be output.

Upon reaching the detector, the ion current is usually amplified and then converted into a digital signal. Most mass spectrometers use electron multipliers as detectors. For the digitisation analogue to digital (ADC) converters and in rare cases, such as in some Waters instruments time to digital converters (TDC) are used. The electronics then convert the continuous ion count into discrete mass spectra for each scan.

A2.2.2.4. Fragmentation and hybrid instruments: The QQQ and Q-ToF

When trying to annotate or even elucidate molecules by means of mass spectrometry, high resolution m/z ratios and isotope pattern provide a good starting point - at least for the determination of the molecular composition [320]. However, to obtain structural information, fragmentation studies are needed. When a molecule breaks up into its part, the resulting fragmentation pattern can be used to match against a database, or confirm the presence/absence of particular chemical groups and their connection. Fragmentation of molecules can occur in various ways; unintended or planned. Unintended fragmentation, for example the dissociation of labile molecules in the source may yield structural information, but is mostly undesired - particularly in metabolite profiling studies. Intended fragmentation can be achieved by the source design, for example by electron impact ioni-

sation in GC-MS, or by collision-induced dissociation (CID) in LC-MS. CID fragments by accelerating ions to provide them with kinetic energy and then collide them with neutral molecules, such as helium, nitrogen or argon gas. The kinetic energy, adjustable by the accelerating voltage allows to control the intensity of the collision and therefore breakage of particular bonds. Different voltages will result in different fragmentation patterns of the same molecule.

CID is mostly performed on two types of instruments. The triple quad (QQQ) and the Q-ToF.

The triple quadrupole, first proposed by Yost *et al.*[656], holds three quadrupole mass analysers in series, with the middle, quadrupole either operating as regular ion guide or collision cell (filled with collision gas). When the second quadrupole (Q_2) is operated as collision cell, four basic types of fragmentation experiments can be performed [129]. A product ion scan (Q_1 selects m/z , Q_3 scans all m/z), a precursor ion scan (Q_1 scans all m/z , Q_3 selects m/z), neutral loss scans (Q_1 and Q_3 scan all m/z with constant offset) and selected reaction monitoring (SRM) (Q_1 and Q_3 select single m/z). Particularly SRM, or its extension to multiple masses at a time, multiple reaction monitoring (MRM), has become indispensable for targeted LC-MS profiling methods [29, 68]. This is, because compared to untargeted profiling by a quadrupole mass analyser, SRM methods have a very low LOD, as the mass analyser can dwell only on one or few ions at a time. QQQ methods are also the method of choice for quantification by mass spectrometry.

A major disadvantage of QQQ however is their inability to perform high mass accuracy measurements - a particular drawback for untargeted profiling. The Q-ToF, however, is a high mass resolution instrument, with CID capabilities. It has a quadrupole for mass selection, a CID chamber and an orthogonal ToF for detection, as first proposed by Morris *et al.*[421, 422] and illustrated in figure A2.16. As it allows for ToF-like high mass accuracy untargeted profiling, but also for targeted MRM-like methods, it has become one of the most adopted instruments in the metabolomics community. Many LC-MS experiments of this thesis will be performed by Q-ToF mass spectrometers.

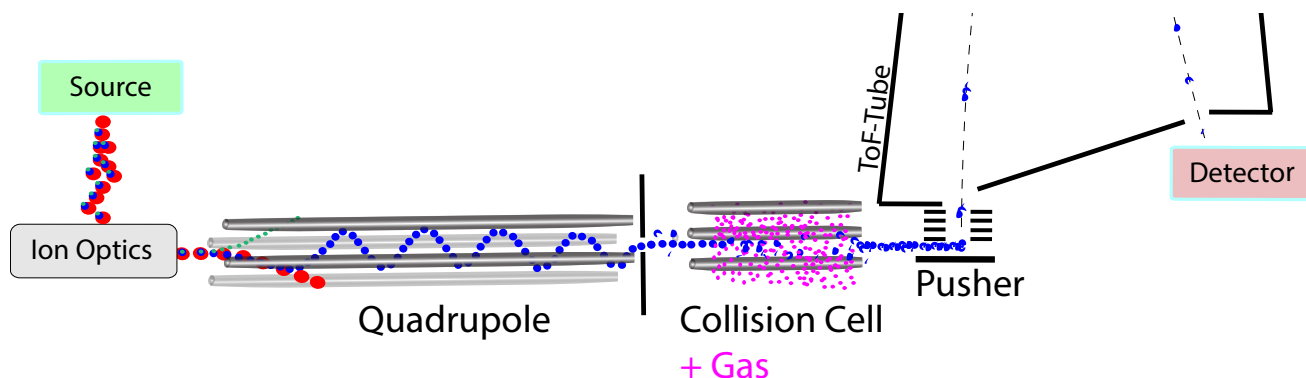


Figure A2.16.: **Q-ToF (quadrupole time-of-flight) mass analyser:**

The Q-ToF combines the quadrupole unit from figure A2.14 and the time of flight module from figure A2.15, separated by a collision cell. The quadrupole allows to select a m/z and fragment it by collision with inert gas in the collision cell. Diagnostic fragments can then be detected using the ToF mass analyser.

A2.3. Hyphenated approaches

Even though NMR spectroscopy and mass spectrometry are very powerful analytical methods in their own right, their full potential may be unleashed when online-coupled to a prior separation stage; particularly for complex mixture analysis, such as performed during metabolomics experiments. However, coupling analytical methods may create technical problems and also make it necessary to constrain the individual parts.

The following sections will discuss the fundamental principles, potential, limitations and the general workflow for LC-NMR (A2.3.1), LC-MS (A2.3.2) and GC-MS (A2.3.3).

A2.3.1. LC-NMR

Pioneered in the 1970s using low field strength iron magnets, LC-NMR has become a mature laboratory technique. It is primarily used for complex mixture analysis, such as pharmacokinetic or natural product studies, when structural information, for subsequent structural elucidation is desired [157].

LC-NMR can be performed in various modes. On-flow, stop-flow and with solid phase extraction SPE (cartridges). On-flow NMR online-couples the eluent flow of a HPLC separation to a NMR magnet, as illustrated in figure A2.17. A major disadvantage, is the short

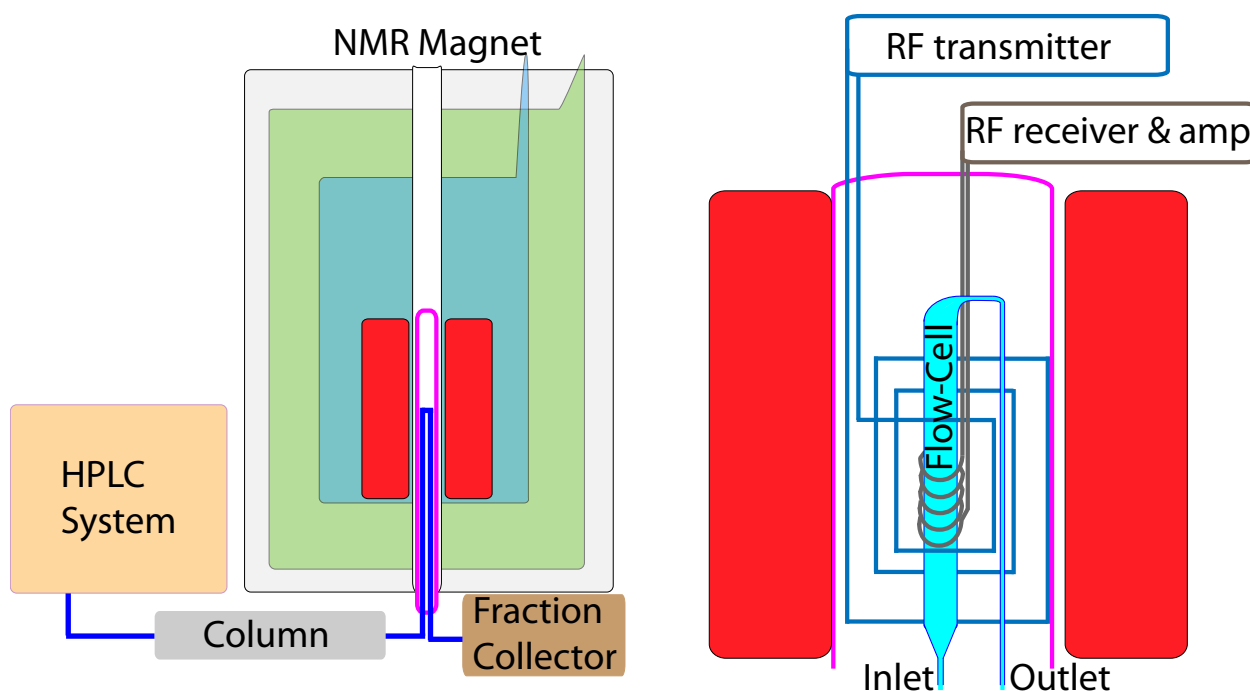


Figure A2.17.: **On-flow LC-NMR system:**

An HPLC system is coupled on-line to a NMR magnet. Fraction collection after NMR acquisition is optional. The right side shows a magnification of the flow-probe. The eluent with the analytes is continuously pumped through a flow cell, instead of a regular 5 mm tube. [6].

residence time of a peak inside the flow-cell, which results in a only low number of scans to be averaged, and therefore a bad SNR. Stop-flow uses the same setup. However, as soon as a peak of interest reaches the flow-cell, the flow is stopped and an arbitrary number of scans and experiments can be performed on the trapped peak. If the flow is stopped for too long, diffusion will compromise the chromatographic performance. In the SPE cartridge approach an offline separation is performed and the individual fractions dried. The individual fractions are then subsequently eluted into the flow-cell using deuterated solvents. The other two approaches require expensive deuterium oxide and deuterated d_3 -acetonitrile as solvents.

Even though solvents are usually deuterated, solvent suppression for the residual water and H_3 -acetonitrile is required. However, as the water resonance frequency depends on other solvents in the mixture (acetonitrile), it will shift throughout the HPLC gradient [305] and suppression of a single frequency is of little use. Thus, a modern LC-NMR spectrometer ex-

ecutes a survey scan prior to the acquisition scan. The survey scan registers the new solvent resonance frequency, which can then be suppressed during the subsequent acquisition scan.

For the separation in LC-NMR a reversed phase gradient of D_2O to d_3 -acetonitrile is typically used [535, 207]. However, a few other types of chromatography, such as HILIC [201], ion exchange [133] or ion pairing [525] have been reported.

Chapter III.3 will use a LC-NMR system, to compare the performance of a reversed phase and an ion pairing method, particularly adapted for LC-NMR. The on-flow results are then compared to a manual HPLC fractionation, analysed by conventional tube NMR.

LC-NMR is also frequently coupled to a mass spectrometer, to record m/z and even fragment information, to aid the structure annotation/elucidation process. LC-NMR-MS analysis splits the flow in a split ratio of 90 % to 10 % for NMR and MS, respectively. The MS eluent may be supplemented with a small amount of organic solvent to aid ionisation [111].

A2.3.2. Liquid chromatography coupled to mass spectrometry (LC-MS)

A liquid chromatography system, as described in section A2.1.2.1 can be coupled online to a mass spectrometer (section A2.2.2), by infusing the chromatographic eluent into the source of the mass spectrometer. The source is needed to evaporate the solvent and ionise the analytes. There are different types of source that would be compatible with a LC-MS system. However, the most prominent LC-MS source by far is the electrospray ionisation (ESI) interface, developed by John Fenn [170].

A2.3.2.1. Electrospray ionisation

Electrospray is an atmospheric pressure ionisation mechanism. Atmospheric pressure ionisation mechanisms are very sensitive [65], soft (little in-source fragmentation), and allow simple sample introduction and coupling of an online separation stage. Their technical challenge relies in gradually increasing vacuum conditions, up to 10^{-8} mbar at the detector, whilst still allowing sufficient ions to pass into the instrument. Figure A2.18 shows the assembly of an ESI source and the ESI mechanism. The vacuum is built up over three stages, the ionisation, the focusing and the detector stage, with the first served by a rotary

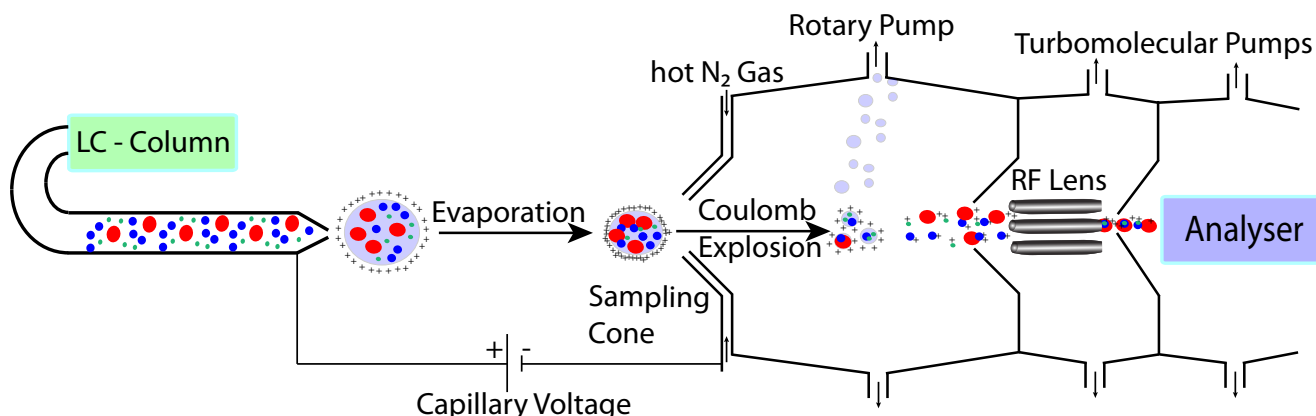


Figure A2.18.: **Electrospray ionisation (ESI) mechanism:**

The illustration shows the working mechanism of an electrospray ionisation source, operated in positive ionisation mode. The size of the individual components, particularly the source to droplet ratios are not authentic, as this drawing is for illustration purposes of the mechanism only [88].

pump and the last two by turbo molecular pumps. In the first step of the ESI mechanism, the LC eluent is dispersed into small droplets by passing through a charged steel capillary. All droplets are accelerated towards the first vacuum stage by a high electric field, applied between the capillary and the sampling cone (3-6 kV). The direction of the field determines whether the source operates in positive (ESI+) or negative (ESI-) mode to generate cations or anions, respectively. Before entering the first vacuum stage, the droplets need to pass a curtain of hot, inert gas (N_2), which causes them to evaporate - aided by the vacuum. This causes the solvent droplet, with the dissolved analytes, to shrink. The charges, created by removing or adding electrons through the capillary voltage, sit on the surface of the droplet. However, with a shrinking diameter, but constant electron excess/shortage, those charges move closer together. At a certain droplet size, the repelling forces between the opposing forces become too high and the droplet is torn into pieces. This process, called Coulomb explosion, disperses the droplet into smaller droplets, which are subject to further evaporation. This whole process repeats until all solvent has been removed by the vacuum and the "naked ions" remain. Another voltage accelerates the ions of the desired charge into the second vacuum change. All non-charged particles are removed by the vacuum. At the second stage, the vacuum increases and a RF lens focuses the ions into a tight beam, which is passed into the third stage - the detector stage.

At this stage only ions of the desired charge are left [49, 88, 445, 304, 334].

In the simplest case, the ion of a neutral molecules (M) is formed by adding or subtracting a proton (M^+ ; M^-) in positive and negative mode, respectively. Therefore the peak in the mass spectrum will be increased or decreased by the weight of one proton ($M^+ = M + 1.00727$; $M^- = M - 1.00727$). The molecule hydrogen pair is called adduct. In a LC-MS run, many other adducts, apart from the (M^+ ; M^-) are likely to be found [574]. In positive mode, sodium ($M^+ + Na$) and potassium ($M^+ + K$) are frequently observed. In negative mode, the formation of formic acid adducts ($M^- + FA$) are common, if formic acid was added to aid ionisation [641] or improve chromatography. It is also possible for molecules to cluster during adduct formation ($2M^+ + H$). Electrospray is also known for producing multiply charged ions, particularly for molecules above 1 kDa. This is very helpful for protein analysis as the m/z ratio of multiply charged molecules, that otherwise would exceed the range of the detector, can easily be measured. Adducts and multiple charges need to be taken into account when performing database searches.

Coupling two analytical methods, may also lead to constraints of the individual ones. So is the case in LC-ESI-MS. Buffered solvents, such as phosphate buffers, are commonly used for LC-UV applications. However, the ESI interface is incompatible with high salt concentrations. In the simplest case, those may lead to heavy buffer-adduct formation. Furthermore, they will contaminate the source, by building up salt crystals around the cones, compromising their diameter and leading to a sensitivity drop throughout the run. It may also happen, that high salt concentrations suppress the signal completely, in a process called ion suppression. An exception are volatile buffers, such as formic acid, which are often added to ensure chromatographic reproducibility and aid ionisation [641].

Ion suppression is the suppression of an analyte signal by another compound inside the system [18]. Research suggest that less volatile molecules adversely affect droplet formation and solvent evaporation, when in high concentration [17]. Those non volatile molecules may be salts, ion pairing agents, but also highly concentrated endogenous compounds. They lead to reduced sensitivity for other analytes or even quantification errors. As one of the dogmas of metabolomics is that a single compound's signal intensity is proportional to its concentration, when compared across several samples, ion suppression can be a major pitfall. Thus it is to be avoided as much as possible. In order to reduce ion suppression

the use of high quality solvents is essential. Sample cleanup, often as part of the sample extraction, may help [388]. However, most importantly, a good chromatographic separation, minimising the number of analytes eluting at the same time, is mission critical.

A2.3.2.2. LC-MS sample set acquisitions

This section describes how a typical LC-MS profiling experiment, such as performed for the mutant screen in chapter V. is set up.

As for each platform, the first step is sample preparation. Modern LC systems do not tolerate any non-soluble particles. Therefore those need to be removed by cold centrifugation, after reconstitution with the starting condition solvent. Supernatants can be transferred into a 96 well plate as was the case for the extraction comparison study in chapter II.. Individual high recovery vials, such as for the mutant screen and isotope labelling experiments in chapters V. and IV., respectively, may also be used. Another important step is checking the condition of the LC-MS system. Particularly as the used UPLC systems operates at pressures up to 1000 bar, it is important to check for improperly mounted fittings or leakages. It is also important to clean the electrospray source prior to use. As mentioned above, salts and molecules can deposit onto the surfaces, particularly onto the cones close to the orifice and compromise the overall LOD. After the source has been resembled, the mass spectrometer's mass accuracy may be calibrated to achieve as low as possible ppm errors. This is usually performed by infusing a mixture of compounds with known m/z ratios. In this thesis sodium-formate, which forms specific clusters across the whole m/z range, is used. Furthermore all voltages, such as capillary, extraction or detector voltage are adjusted to achieve maximum sensitivity. At this stage, the latest, the solvents may be topped up or swapped. Before starting the actual experiments, several blanks and quality control (QC) samples may be injected. Blanks are helpful to estimate background arising from the system or solvents. High background signal may not only lead to false positively detected peaks, but is also a major risk factor for ion suppression [17]. QC samples, which are a representative pooled mix of sample aliquots, are used to monitor the system's stability throughout the run [511]. However, they are also well suited to condition the chromatography system. During the conditioning phase, the QC mix is injected repeatedly, until the chromatographic system is stable, *i.e.* the retention time drift

between injections has become negligible and chromatograms overlay nicely. It is also recommended to check the correct operation of the lock spray, to ensure high mass accuracy. If the whole system performance lies within the expected quality standards, acquisition of the sample set may commence [616, 141, 613]. However, it may be necessary to use intensity correction models for high sample number acquisitions, as described by Dunn *et al.*[141].

A2.3.2.3. Different types of chromatograms: TIC, BIP, EIC

After, or also throughout acquisition, data must be analysed. As LC-MS data is three dimensional, visualisation is not straightforward. Even though data can be plotted three dimensionally, for example by using coloured heat maps, this is not often done. Most times, the 3D LC-MS data is plotted as a 2D projection, with the m/z domain collapsed. This yields a relative intensity versus retention time plots. However, as abolishing the m/z dimension leads to a loss in information, three types of chromatograms, collapsing the m/z domain in different ways, exist. Figure A2.19 shows examples of those.

Total ion chromatogram (TIC): The TIC sums all m/z intensity values of a scan. It therefore corresponds the total ion current of a scan. As can be seen in figure A2.19a, the baseline drift is clearly visible. Thus a TIC is useful to check peaks with respect to background (noise).

Base ion peak chromatogram (BIP): The BIP only displays the intensity of the most intense peak at each scan. Therefore the BIP is dominated by the most intense peaks (figure A2.19b). Compared to a TIC, the baseline is not considered. A BIP can be useful to find peaks, that can be further isolated in an EIC.

Extracted ion chromatogram (EIC): The EIC is the chromatogram of a single m/z slice of defined width, such as all ions between 338.2 and 338.4 Da in figure A2.19c. Only one peak seems to lie between those two masses. Therefore it is well suited to display (the chromatographic profile of) individual peaks. Figure A2.20 shows the mass spectrum of the peak at 15 min.

A2.3.3. Gas chromatography coupled to mass spectrometry (GC-MS)

Gas chromatographs were first coupled to mass spectrometers, decades before LC-MS was invented. McLafferty [402] and Gohlke [202] were the first to couple a quadrupole and a

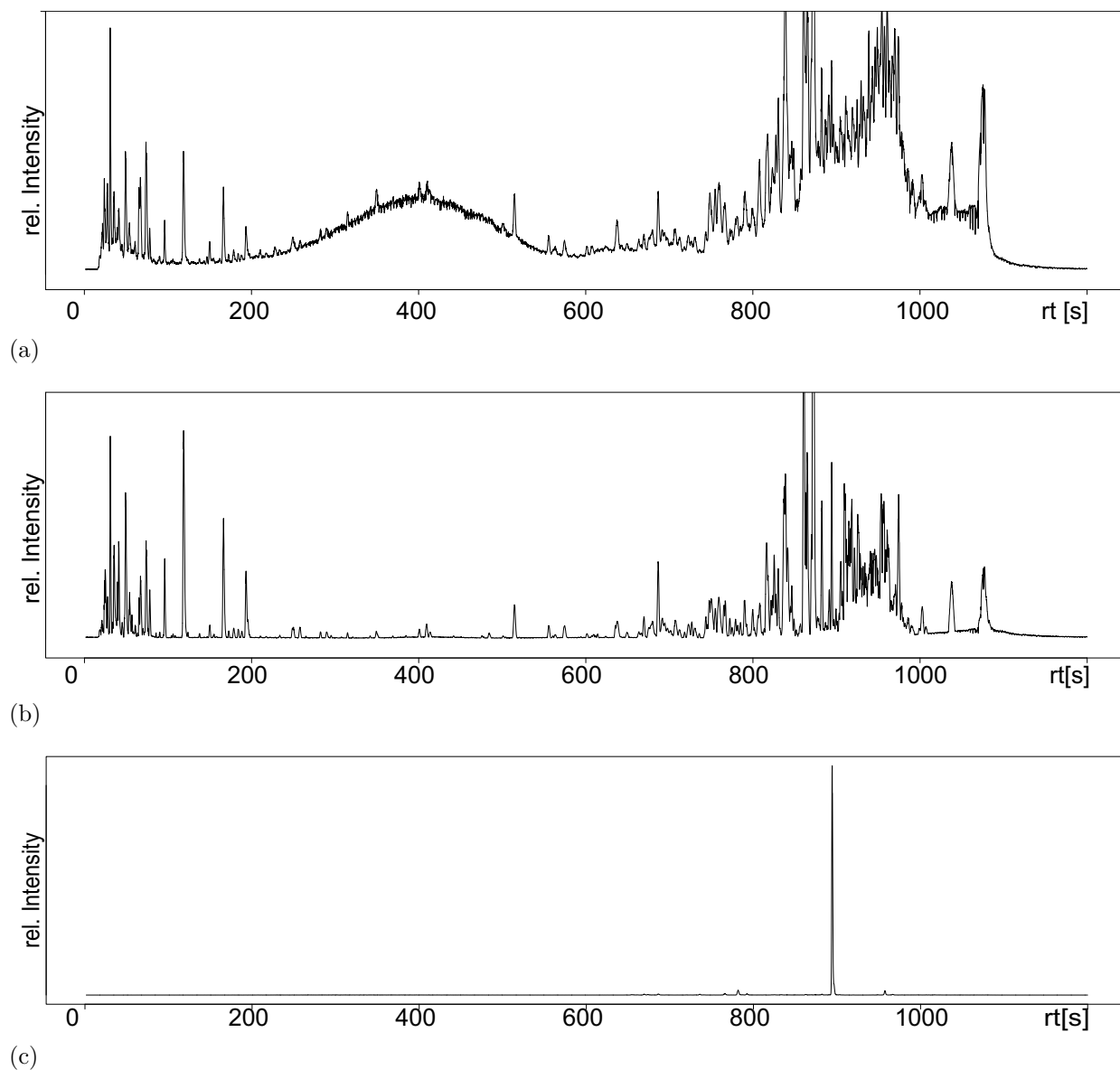


Figure A2.19.: **Different types of chromatograms (TIC,BIP,EIC):**

Total ion (TIC) (figure A2.19a), Base ion peak (BIP) (figure A2.19b) and a $m/z = 338.3$ Da extracted ion (EIC) (figure A2.19c) chromatogram of a *C. elegans* extract acquired in positive ESI mode.

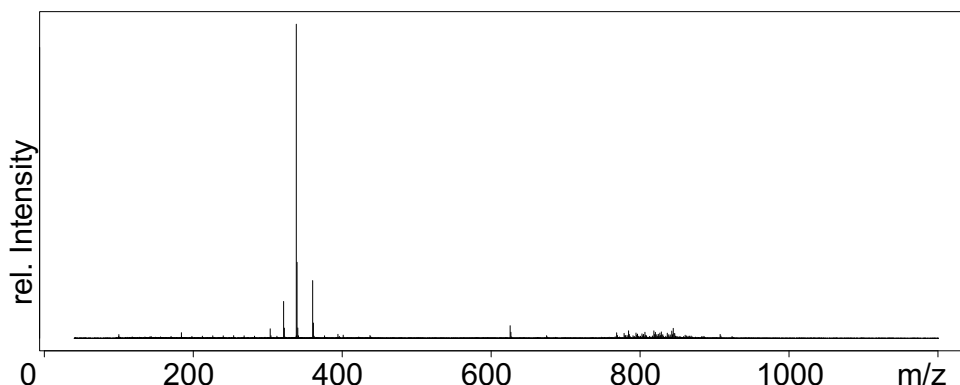


Figure A2.20.: Mass spectrum of peak apex from EIC in figure A2.19c:

ToF mass analyser, respectively, to a gas chromatograph. This has become possible after the invention of the electron impact ionisation source, by J. Benyon [46], who realised the potential of fragmentation patterns to aid the identification of chemical compounds. The combination of machine independent retention indices and diagnostic fragmentation patterns, generated by each molecule, make GC-MS one of today's most important analytical platforms.

A2.3.3.1. Electron impact ionisation (EI)

Electron impact ionisation (EI) is a "hard" ionisation technique, breaking up the molecule into diagnostic fragment ions. A typical EI source is illustrated in figure A2.21.

In the EI source, the non-charged molecules, arriving from the GC-column are bombarded with a stream of electrons. Those are generated by a light bulb like cathode - the filament, and accelerated towards the anode. The electrons do not impact the molecule, but rather interact with its molecular bonds, which leads to excitation and breakup of the bond [42]. By convention, an electron collision energy of 70 eV is used, to make spectra comparable across different machines. Indeed, EI spectra are very reproducible, allowing database matching (section A3.1.4.1). After fragmentation, the ions are then accelerated between the electric field of the repeller and extractions lens, and separated by they m/z ratio prior to detection.

Figure A2.22 shows the EI spectra of caffeine (figure A2.22a) and malic acid (figure A2.22c),

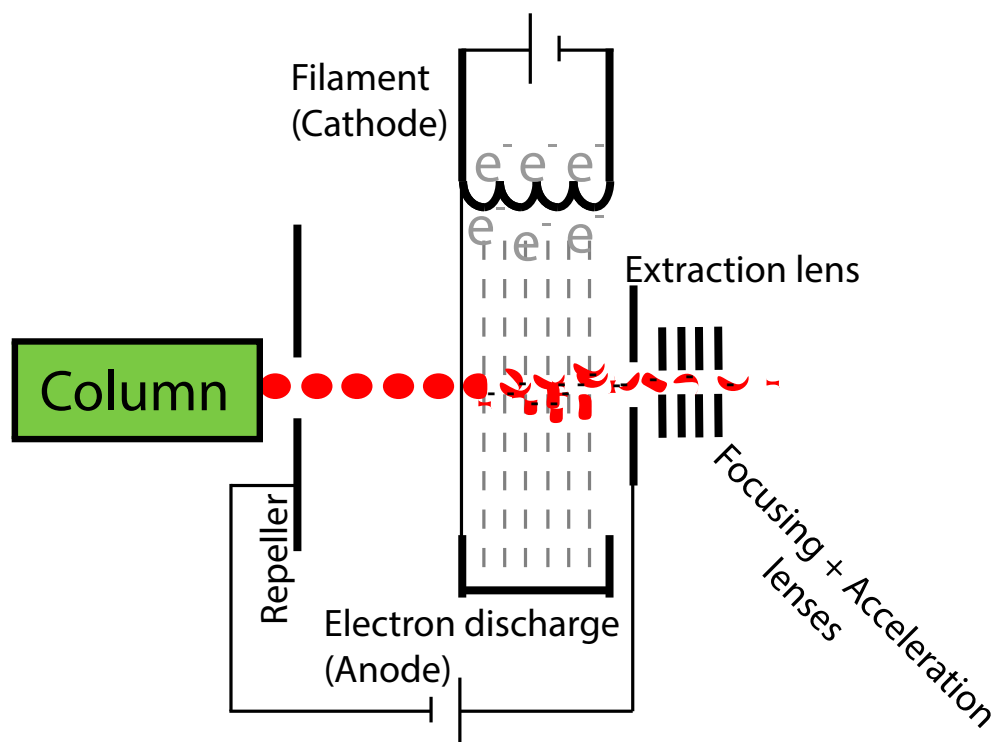


Figure A2.21.: **Electron impact ionisation (EI) source:**

Molecules are fragmented and ionised by electron bombardment.

together with a fragmentation tree for caffeine (figure A2.22a).¹

For caffeine, all major fragment ion can be explained by specific fragments. This shows the strength of EI-GC and fragmentation studies in general, to assign and elucidate molecules. Fragmentation occurs according to complex rules, such as rearrangements [403]. Fragments may not only be detectable ions, as figure A2.22a shows. Also radicals, such as $\cdot CH_3$, or neutral, gaseous molecules, such as CO , referred to as neutral losses, are common fragmentation products.

When using derivatisation agents, it is possible, that the major fragments are due to fragmentation of the groups added by the derivatisation agent. The MSTFA derivative of malic acids displays such a case (figure A2.22c). Still, some diagnostic fragments may be found in lower abundance. However, if those, in conjunction with retention time, do not

¹All fragment ions are putative, but can then be tested by accurate mass measurement and isotopic labelling studies.

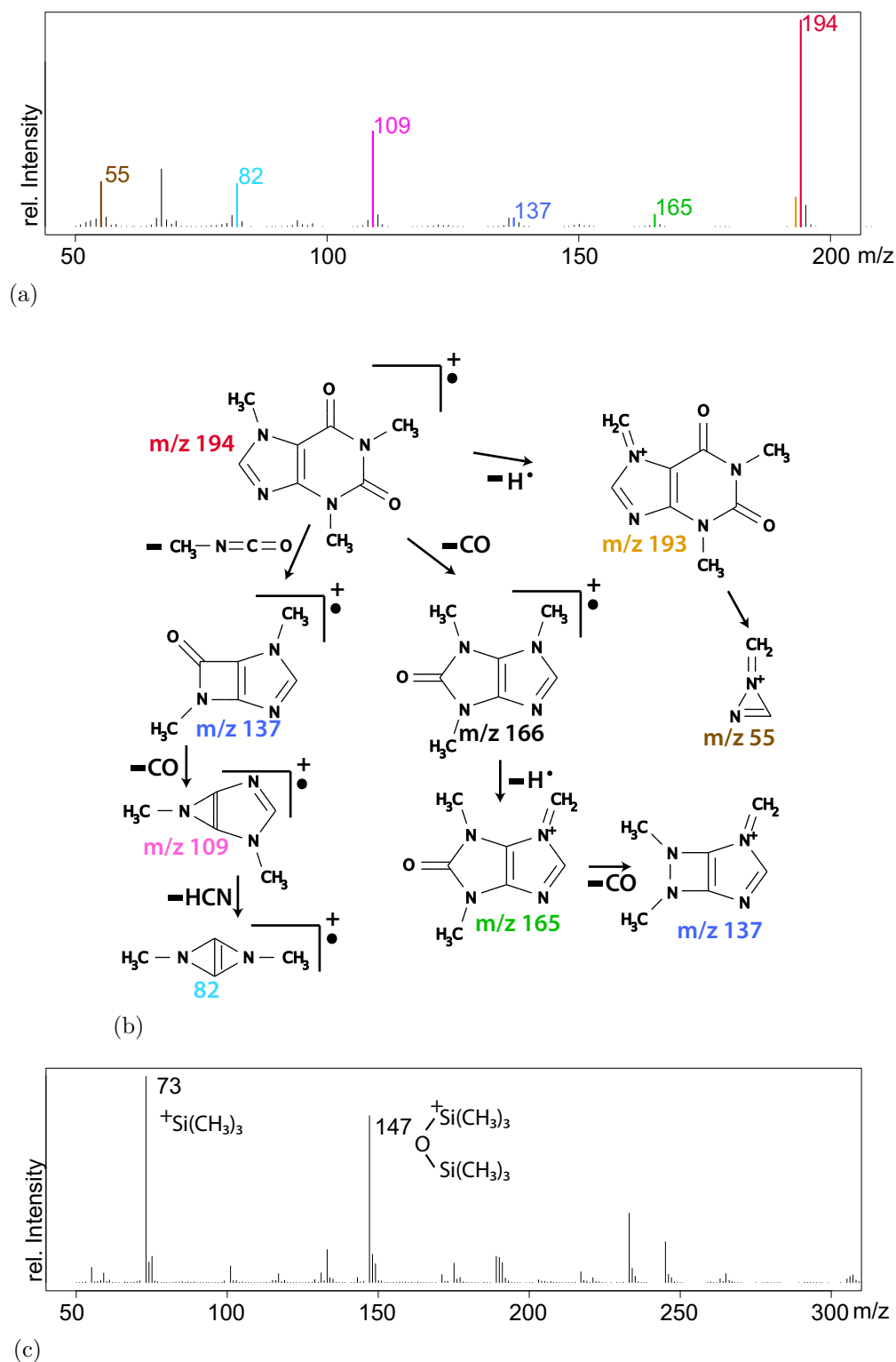


Figure A2.22.: **GC-EI-MS fragmentation:**

Figure A2.22a illustrates the electron impact ionisation (EI) spectrum of caffeine. Major fragment ions are explained by the fragmentation tree in figure A2.22b. A molecule can fragment in different ways, by either a radical or neutral loss. The two most abundant fragments of the EI spectrum of malic acid (figure A2.22c) originate from the MSTFA derivatisation agent, rather than the compound itself.

allow unambiguous assignment of the molecule, the projects in this thesis will refrain from doing so.

A2.3.3.2. GC-MS sample set acquisitions

This section outlines the analytical workflow of a typical GC-MS experiment setup and acquisition, as already done in sections A2.2.1.7 and A2.3.2.2 for NMR and LC-MS, respectively. Figure A2.23 outlines all necessary steps.

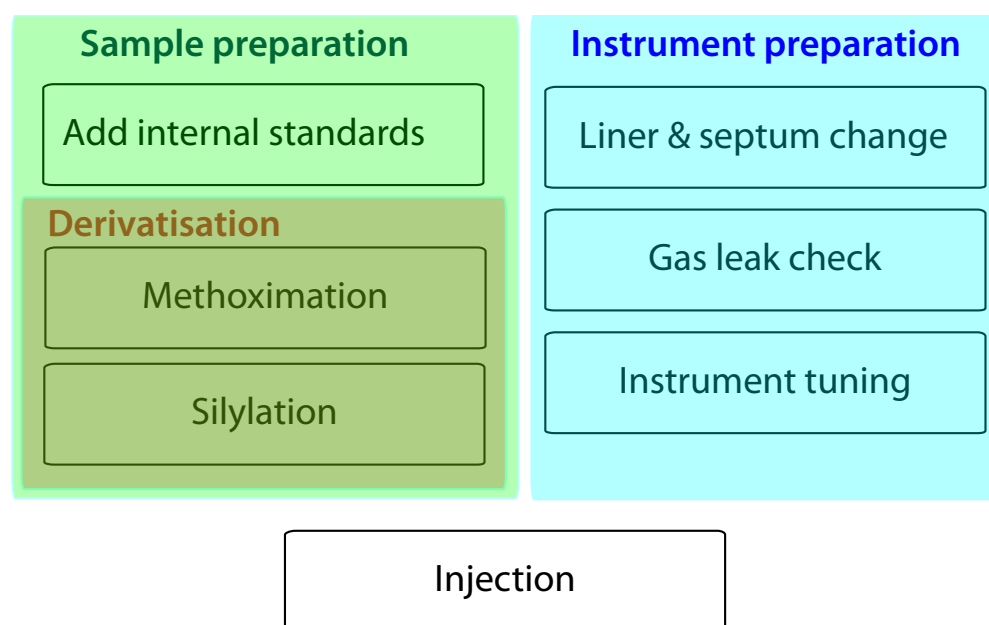


Figure A2.23.: **Standard procedure of a GC-MS sample acquisition:**

Flow chart of a standard procedure for preparation and acquisition of a GC-MS sample set, such as used in this thesis. For details please refer to the text.

In GC-MS the experimental setup divides into instrument and sample preparation, which out of the used analytical methods is the most time-consuming for GC-MS. To the dried samples, the retention time lock standard, d_{27} -myristic acid in this thesis, and quantitation standards, such as $U\text{-}^{13}\text{C}$ -glucose and $2,3,3\text{-}d_3$ -leucine are added. As reproducible derivatisation requires anhydrous conditions, the sample must be thoroughly dried, before incubation with the methoximation solution. After silylation with MSTFA, an injection standard, 2-fluorobiphenyl, is added to correct for small variations in injection volume.

The samples, if not already done so, are transferred into glass vials with inert, silanised inserts.

To prepare the instrument, the liner and septum are changed for each new sample set, as those become contaminated throughout injections. After changing them, it is essential to check for gas leaks. This can be done by checking that the water and oxygen peaks, introduced by ambient air, are below a manufacturer defined threshold. Gas leaks would lead to uncontrollable retention time drifts and oxygen might even damage the stationary phase at high temperatures. When an air-tight system is guaranteed, the instruments can be tuned, *i.e.* lens and detector voltages optimised. Usually performed automatically, the system uses an internal standard to optimise sensitivity, isotope ratios and mass resolution. If all checks are passed and samples are finished derivatising, injections may commence. It is recommended to start of with blank injections and then QC samples to check the system cleanliness and performance, such as retention time reproducibility, sensitivity and peak widths, respectively. [323, 186]

A2.4. How to choose an analytical platform

After all analytical platforms (NMR, LC-MS and GC-MS) used in this thesis have been discussed, an important question emerges: Which platform should be chosen for which application.

Figure A2.24 tries to outline considerations that must be taken into account when choosing the analytical platform for a metabolite profiling experiment.

Nuclear magnetic resonance (NMR) spectroscopy is relatively insensitive and therefore restricted to the analysis of high concentration compounds. Primarily, those are molecules from the central carbon metabolism, such as glucose, amino acids and nucleotides. As it only requires compounds to have protons in order to be visible and does not perform any prior separation, it has no bias towards specific compound classes, as LC-MS and GC-MS would have. However, the lack of a prior separation stage also leads to signal overlap and potential masking of smaller concentrated or similar compounds. Chapter III.3 aims at addressing this problem, by performing a LC-MS analysis of a *C. elegans* extract. Despite the sensitivity and overlap drawbacks, NMR, particularly as assignment is straight

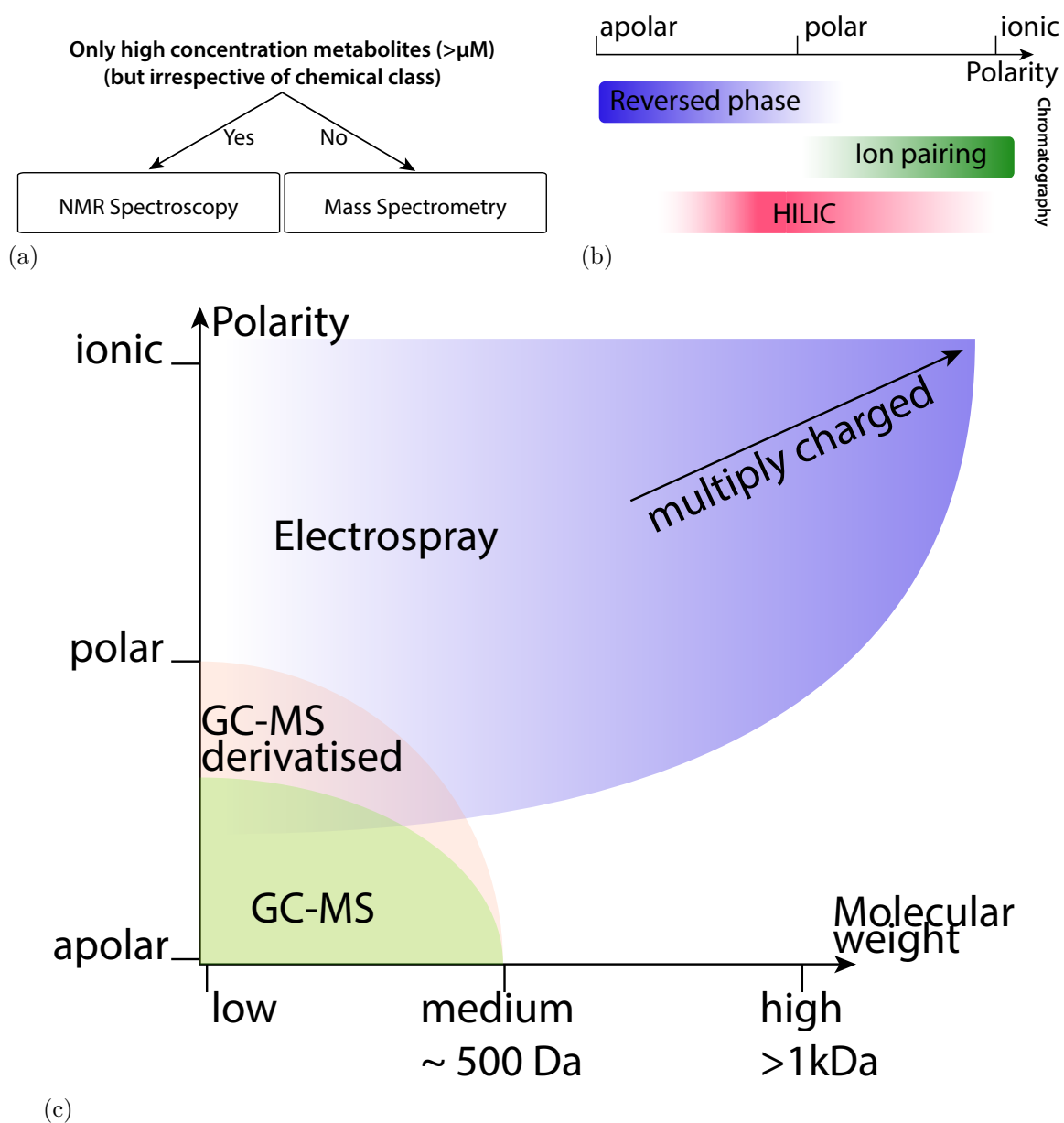


Figure A2.24.: **Choosing an appropriate analytical platform:**

Considerations regarding choice of the analytical platform. Details are discussed in the text.

forward, is well suited for an initial survey.

Mass spectrometry is a very sensitive technique; down to femtomolar concentrations. However, as figure A2.24b and A2.24c show, different types of chromatography and the choice of a certain ionisation interface make it selective towards specific compound classes. GC-MS is very well suited for small and apolar molecules, as it requires them to be volatile, which is inversely correlated with molecular weight and polarity. As the electron impact (EI) source creates ions by fragmentation, the compounds, compared to electrospray, may even be completely apolar, without any ionisable functional groups. GC-MS coverage may be increased by means of derivatisation, *i.e.* the conversion of polar functional groups into apolar ones. However, derivatisation is not an necessarily option for larger polar molecules, as the derivatisation agent (used in this thesis) adds significant weight to the compound and therefore increases its volatility. Electrospray (ESI) is well suited for polar and ionic compounds. It is compatible with a large range of molecular weight. Very low weight molecules, such as small alcohols and acids, may be subject to evaporation and removal by the vacuum pump. On the other end, the formation of multiply charged ions, allows very large compounds to be detected in the m/z (mass to charge ratio) operation range of common mass analysers, such as ToF instruments. Two major drawbacks of LC-ESI-MS are potential problems with ion suppression and difficulty with compound annotation. However, the later will be addressed in chapter IV..

When deciding to use ESI, the selectivity is already determined by the used chromatographic mode, as figure A2.24b shows. Reversed phase is well suited for apolar and medium polarity, such as lipids or vitamins. The more ionic or ionisable groups, the better it is retained by ion pair chromatography. Ion pairs can separate extremely polar molecules, such as multiply charged, highly phosphorylated nucleotides [366]. HILIC is very well suited for polar molecules, but may be extended to cover certain lipid classes [666]. In the case of ion pairing and HILIC, different solvents and modifiers are needed to analyse cationic and anionic ions.

It becomes apparent that there is no analytical platform, which may cover all metabolites, but rather that different analytical platforms compliment each other; particularly if true untargeted profiling is desired. Therefore, the metabolomics community has started to move towards multiple analytical platforms studies. [628, 23, 67, 48, 414, 187, 39, 513].

However, such cross-platform approaches bring challenges on the levels of sample preparation, analysis and data processing and integration. The next section will discuss data processing workflows for the different platforms. The first experimental chapter of this thesis, chapter II., will investigate tissue extraction methods, compatible with all three major platforms; NMR, GC-MS and LC-MS.

A3. Informatics

After all data has been collected, it must be processed and analysed. Particularly for metabolite profiling studies, which accumulate a wealth of analytical data to prise out subtle changes in metabolite concentrations between different samples, computational methods have become indispensable. This section will present the data processing and reduction workflow used in this thesis for each platform. A short introduction into common data analysis, such as data pretreatment methods, univariate and multivariate statistics will conclude this methods chapter.

A3.1. Data processing

Table A3.1 lists the number of data points and resulting file size for individual files, that have resulted from experiments in this thesis.

Data type	Data point / file	File size
1-D NMR	16k - 64k	< 1 MB
2-D NMR	x k by x k (mostly empty)	x - xx MB
GC-EI-MS (low res.)	few ten k	1 – 20 MB
LC-ESI-MS (high res.)	up to 100 M	up to 1 GB

Table A3.1.: ***Typical number of data points and file size of different experiments:***

x and xx are undetermined numbers of one and two digits length, respectively; k and M are the SI prefixes for thousand and million, respectively.

It immediately becomes apparent, that prior to any data analysis and irrespective of the analytical platform, data processing methods need reduce the total amount of information to a few relevant data points (variables). Ideally, they will also remove duplicate

information, such as several spectral variables expressing the concentration of the same metabolite. Furthermore, this condensed information requires alignment across multiple samples, so that the concentration of a particular compound can be compared across all samples. This section discusses the details of these two major steps for NMR (A3.1.1), LC-NMR (A3.1.2), GC-MS (A3.1.4) and LC-MS (A3.1.3).

A3.1.1. Processing of NMR spectra

In a modern NMR instrument, the (90°) excitation pulse consists of a broad range of frequencies, to excite all nuclei of slightly different frequency (different chemical shift) at once. The detected response of such a multi frequency pulse is a complex mixture of decaying harmonic oscillations, of different frequencies, called the free induction decay (FID). In order to yield a frequency spectrum of this time domain signal, fast Fourier transform (FFT) algorithms can be used [110, 153]. However, data processing in NMR may already begin before the FFT. Figure A3.1 outlines all necessary steps from the FID, to an analysable spectrum.

A3.1.1.1. Window functions, phase and baseline correction and shift referencing

Before FFT, the FID may be multiplied by a window function, to modulate the peak shape in the frequency spectrum. For one dimensional ^1H spectra, the most common window function is an exponential decay function, which is multiplied with the FID, as displayed in figure A3.1a. An exponentially decaying function in time domain, corresponds to a Lorentzian filter kernel in the frequency domain. Therefore it is aimed to reduce the noise in the spectrum. The amount of smoothing is determined by the line broadening factor, measured in Hz. If set to high, it may remove diagnostic fine structure of the peaks. Another disadvantage is that with a higher line broadening factor peaks become broader. After FFT, the peaks must be "phased". Usually, the FID oscillation, detected in the M_x axis is not a perfect cosine wave, *i.e.* it's maximum is not at 0 ($\cos(0) = 1$), but rather starts later, *i.e.* it has a phase shift. Therefore after FFT its frequency spectrum is distorted as shown in figure A3.1b. However, by adjusting the phase angle to transform the FID into a "perfect" cosine wave, an aesthetically pleasing, interpretable NMR spectrum is yielded. The next steps are baseline correction (figure A3.1c) and referencing the ppm scale (figure A3.1d). For baseline corrections different methods exist [644]. In most cases

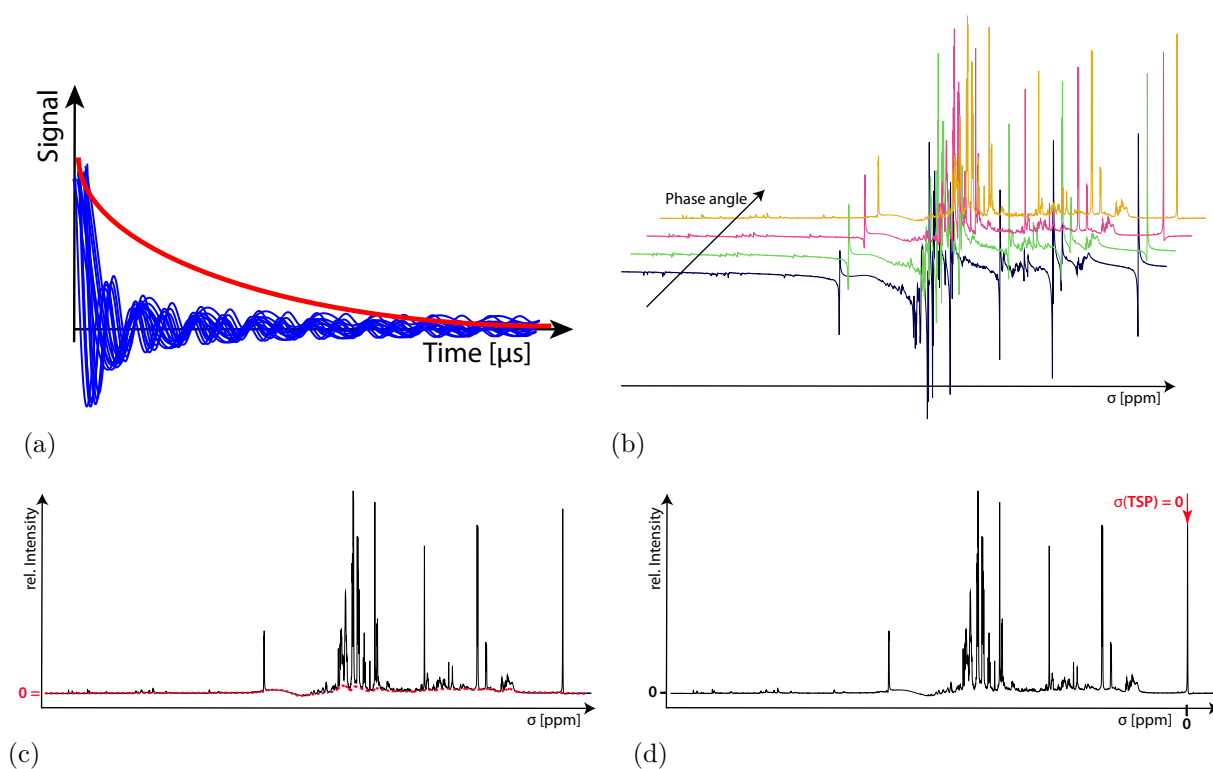


Figure A3.1.: NMR spectra processing workflow:

Figure A3.1a shows an exponentially decaying window function, to be multiplied with the FID prior to FFT to smooth the peaks. The next inevitable step, is to correct the phase angles of the frequency spectrum (figure A3.1b). If desired, the baseline (figure A3.1c) and chemical shift (figure A3.1d) offset can be corrected. For detailed explanations, please refer to the text.

a polynomial or spline function are fitted. For the referencing, either the added internal chemical shift reference, such as TSP or a metabolite of known shift, that is unaffected by pH variation, are used.

A3.1.1.2. Alignment versus fitting or binning

After the individual spectra have been processed, they can be compared. In the easiest case by spectral overlay as done in figure A3.1b. To automatically compare and analyse spectra, for example by multivariate models, they require further processing, due to small variations in chemical shift. Those small variations are introduced by difference in pH, or concentrations in electrolytes, metabolites or proteins. The easiest way to overcome this problem is binning. Binning reduces the spectral resolution by dividing it into evenly spaced, user defined or via means of an algorithm calculated intervals [16]. The used intervals should be wide enough to overcome the small variations in peak positions. However, in samples with a different baseline, wide shifts or even peak overlaps, binning is not very reliable and discouraged [115]. Another method is to align the peaks to a prior defined consensus spectrum. Different algorithms are available. In this thesis a recursive segment-wise peak alignment algorithm proposed by Veselkov *et al.*[598] will be used to align the raw data. Recently, peak alignment procedures have also been put under scrutiny, as they, themselves may introduce artefacts. They also fail in the rare case of two peaks crossing their position. In such cases, peak fitting is the best choice. Liebeke *et al.*[365] discuss this problem. Peak fitting tries to quantify a peak by fitting a peak shaped model to its area. This can be done with manual guidance, for example with the Chenomx software suite [617]. Chenomx will be used to quantify the metabolites in the extraction comparison chapter (II.) of this thesis. Recently, Hao *et al.*[229] have released a completely automated peak fitting suite, which has been shown to also work well in the case of large pH shifts, overlapping and crossing peaks [364]. A major advantage of fitting, regardless of manual or automatic execution, is that it ends up with one numeric value for each peak or even compound. This reduces the amount of data, and also removes correlated, redundant variables, which may be advantageous in subsequent modelling procedures. If appropriate internal standards are used, fitted data can also be transformed into fully quantitative data, which is particularly interesting for quantitative pathway modelling tools.

A3.1.1.3. Processing of 2D NMR spectra

The processing of 2D NMR spectra follows the one of 1D spectra. However, depending on the pulse sequence (HSQC, COSY, HMBC, JRES,..) different window functions can improve the results, in particular the SNR. In most cases the FFT will also need to be performed in the second dimension. For phasing and baseline correction, specific peaks across the whole spectrum are chosen. After performing the phasing and baseline corrections on their 1D projections, their consensus correction models are applied to the whole spectrum. Spectra also need to be referenced in both dimensions.

As a COSY spectrum is symmetric with respect to its diagonal, its mirror image can be added to improve the SNR.

A3.1.2. Processing of LC-NMR spectra

LC-NMR spectra may be processed like a stack of 1D spectra. If other detectors are used in parallel or series, retention time alignment between those may become necessary.

A3.1.3. Processing of LC-MS data

As table A3.1 shows, LC-MS acquisitions carry by far the largest amount of recorded information. Also, compared to two dimensional (ppm $\times$ intensity) NMR data, LC-MS files have a three dimensional structure. Each datapoint has an associated retention time (RT), mass to charge ratio (m/z) and intensity value. Also, after reduction, it is required to align the variables of several data files in two, independent (RT $\times m/z$) dimensions, rather than one like NMR (ppm). Thus, they require more sophisticated processing algorithms, to reduce the information. Several software packages and tools are available, which incorporate such algorithms. Most vendors sell their own software suites, for example Waters' MassLynx. Those commercial solutions are usually easy to use black box systems, which are convenient for application focused endusers. However, for explicit method development, also on the data processing side, those are less suited. Therefore, for this thesis, open source software packages were used.

Although many free open source software solutions have been published [604, 547, 277, 618, 409, 315], mzmime [302, 482] and xcms [531, 45] are the most widely used in the metabolomics community. However, all of those work in the same fashion: Peak detection, alignment across files and finding the peaks that were not detected in certain samples in

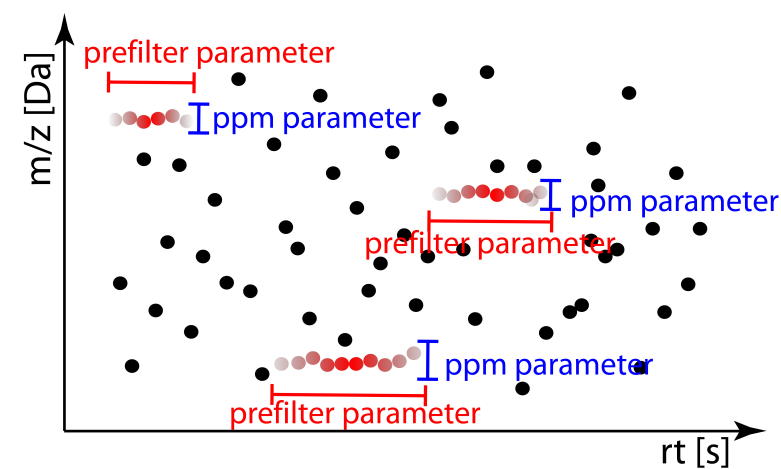
the first place, to yield a peak versus sample intensity matrix. Usually one package offers a set of different algorithms for the same step. Not rarely do different software packages adopt each other's methods, if they were found to perform well. Compared to *mzmine*, which requires its modules to be written in C, *xcms* is implemented inside the open source statistical programming suite R, with the option to write and use C libraries. As R is much better suited for code prototyping, it was decided to base this thesis's experiments and tool development on the *xcms* software package

A3.1.3.1. The LC-MS deconvolution and alignment workflow: XCMS

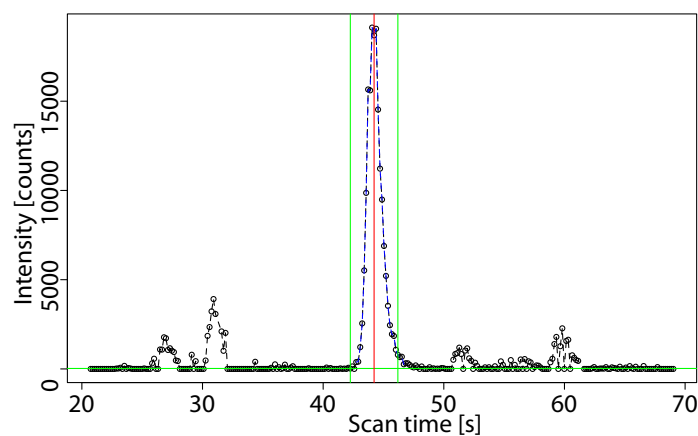
Out of the four experimental chapters of this thesis three apply *xcms*, optimise its parameters or even build tools onto its existing algorithms. Therefore this section describes its workflow and important parameter settings in more detail. Parameters are highlighted in *italics*.

Before data can be loaded into R, it must be converted, from the vendor's proprietary file format, to an open source format, such as *mzXML* [472]. Furthermore profile mode data, *i.e.* mass spectrometry data which also records the peak variance in the m/z domain, could be reduced to centroided data, *i.e.* peaks that are only described by their apex m/z value. After those steps, *xcms* can be launched inside an R session. The first processing step is peak detection, to identify chromatographic peaks within an abundance of noise. Peak detection for high mass resolution data ("centWave" algorithm [566]) is performed in two steps, illustrated by figure A3.2.

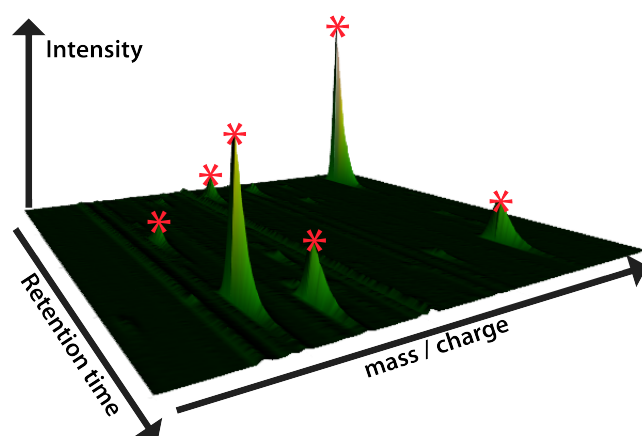
At first regions of interest (ROI) are defined. ROI are traces in the chromatogram, where a signal, above a certain intensity can be seen for a minimum number of consecutive scans, with minimal m/z variation in-between those scans. The number of consecutive scans and minimum intensity are specified by the *prefilter* parameter. The minimum ppm deviation in the m/z domain is defined by the *ppm* parameter. Figure A3.2a gives a graphical representation of this procedure and its parameters. As not all ROI are peaks necessarily, a second, step checks whether the ROI has the typically expected shape of a LC-MS peak. A wavelet ("mexican hat") filter kernel, which corresponds to the second derivative of a



(a)



(b)



(c)

Figure A3.2.: **Wavelet peak detection in xcms:**

First, the regions of interest (ROI) are found (figure A3.2a), which are then fitted with a wavelet filter kernel (figure A3.2b). All peaks which fit the wavelet well are identified as peaks (figure A3.2c).

Gaussian function, is used as model peak shape. The model's width, its wavelet coefficients, are determined automatically. However, it is recommended to use the *peakwidth* parameter to specify the expected minimum and maximum width of a peak expected under the chromatographic conditions. This parameter is particularly important, when UPLC systems are used, as the default parameters are set for the much wider HPLC peaks. The wavelet also determines the peak's apex and integrated intensity, and its retention time and m/z boundaries, as illustrated in figure A3.2b. By knowing a peak's apex intensity, its signal to noise ratio (SNR) can be calculated, by relating it to its chromatographic neighbourhood. Using the *snthresh* parameter, a minimum SNR, below which peaks are discarded, can be set. However, as noted by Dunn *et al.*[142], this SNR cannot necessarily be compared with classical SNR values and therefore settings as low as 1 may yield reproducible peaks. An extra parameter requires to be set, when using Waters instrumentation with lock mass corrections. During acquisition of the mass correction lock spray every few seconds, the eluent spray is interrupted for two scans, leaving gaps in the chromatogram. An algorithm, developed by Benton *et al.*[43] can fill those gaps by interpolation, when the *lockMassFreq* parameter is set to "TRUE". At the end of the peak detection procedure, all chromatographic peaks are identified (figure A3.2c) and their intensity and position is recorded inside a table.

In LC-MS detected peaks are often referred to as "feature". This acknowledges that the origin of the peak is not yet known. The term metabolite or compound would be misleading, as it could also be an isotope peak, adduct or cluster. Thus, the term feature will be used throughout the thesis.

The next step of the *xcms* workflow is to align the features, which originate from the same compound, across multiple files. In *xcms*, this is performed in two steps, which can be repeated iteratively until a satisfactory result is reached. In a first step, coarse RT and m/z windows are spanned to "group" features with a very similar mass and retention time across different samples. The RT and m/z boundaries of these windows can be defined by the *bw* and the *mzwid* parameter of the grouping function, respectively. In the ideal case, the windows are large enough to allow for mass inaccuracies and naturally occurring retention time deviation, but small enough to not pool features of different origin. If for a certain group one and only one feature was assigned from a datafile, this groups is regarded as "well behaved" and most likely to originate from the same compound. In a

second step, those well behaved peak groups are used to estimate the local retention time deviation (induced by variations in RT), for each file. This is done by calculating the RT deviation of each feature from its group's median RT. This RT deviation from a group's median is used on a per file basis to build a model of each file's RT deviation across the whole retention time range, by fitting a LOESS (locally weighted scatterplot smoothing) regression model [105]. This procedure is illustrated by figure A3.3.

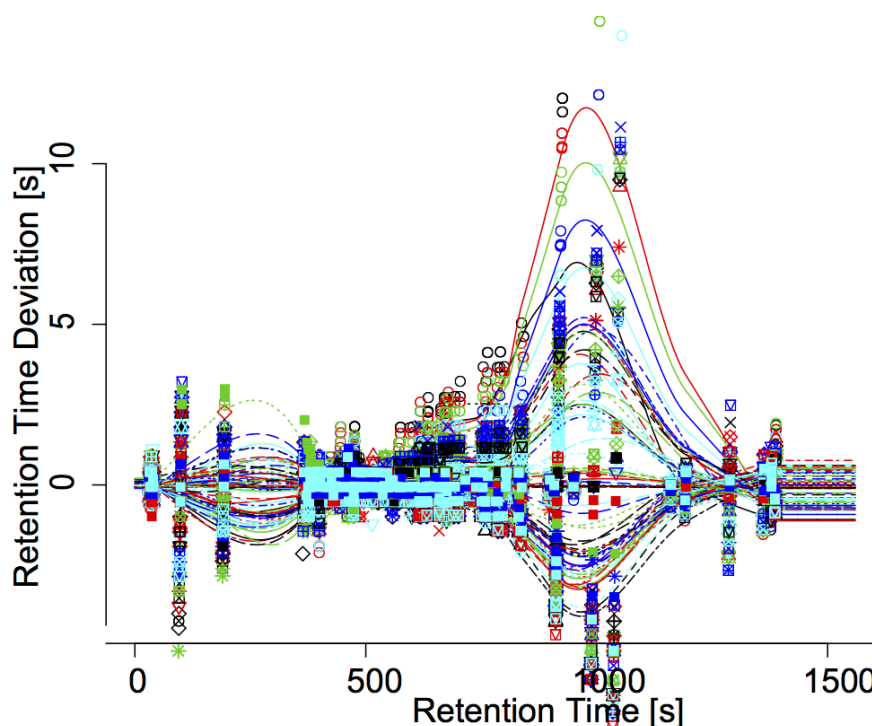


Figure A3.3.: **Retention time alignment in xcms:**

The retention time deviation from the median RT of the "well behaved" peak groups (dots) is modelled across the whole RT range using LOESS regression (solid line). Individual colours correspond to different sample files (acquisitions).

Using this model, all features are aligned to the median RT. In fact, also the times of the individual scans are aligned, allowing to predict the expected RT position of a feature, which was not found in a particular sample. Thus, the correct area (where the feature should have appeared) can be calculated in the final step. For the alignment step, the most important parameter is the *span* parameter. It defines the degree of polynomial smoothing, *i.e.* whether individual input data points may affect the curvature of the resulting model

or the model is more representative for the average of data points. Finding a satisfying *span* parameter requires the user to estimate whether the retention time model, as shown in figure A3.3 is realistic. The grouping and alignment step should be repeated iteratively, to refine the group's window size, which in return will allow a more precise retention time deviation model. It is reasonable to aim at a maximum retention time deviation of less than one second, after correction [44].

At this stage, one already has a feature (variable) by sample (observation) intensity matrix. However, this matrix may still contain missing values, in case a feature was not found during the initial peak detection step. As leaving zeroes would negatively impact any subsequent statistical analysis, those missing values require to be filled. The retention time deviation model can predict the expected position of the missing feature from the not well behaved groups (*i.e.* those features not found in each sample). Thus, in order to complete the intensity matrix, *xcms* integrates the mass spectra at the missing features' expected positions.

Finally, a complete feature by sample matrix can be exported for subsequent data analysis.

A3.1.4. Processing of GC-MS data

Even though compared to LC-ESI-MS, GC-EI-MS files contain far less data points and are recorded in low mass resolution (in this thesis), they are of similar complexity. As illustrated in figure A3.4, a GC-EI-MS data file structure can be viewed as sparsely populated retention time $\times$ integer m/z matrix, with several fragmentation patterns. Even though fragmentation patterns are complex, they are very diagnostic and allow to identify a compound, as discussed in section A2.3.3.1.

A3.1.4.1. Fragmentation patterns and database matching

Indeed, initiated by McLafferty in the 1970ies, extensive fragmentation and retention index databases for GC-EI-MS have been curated over the last few decades [335]. The most comprehensive current library may be the NIST/EPA/NIH Mass Spectral Library [25], with over 200 thousand entries. As it would be impractical to manually compare recorded fragmentation patterns with the database's for identification purposes, efficient matching

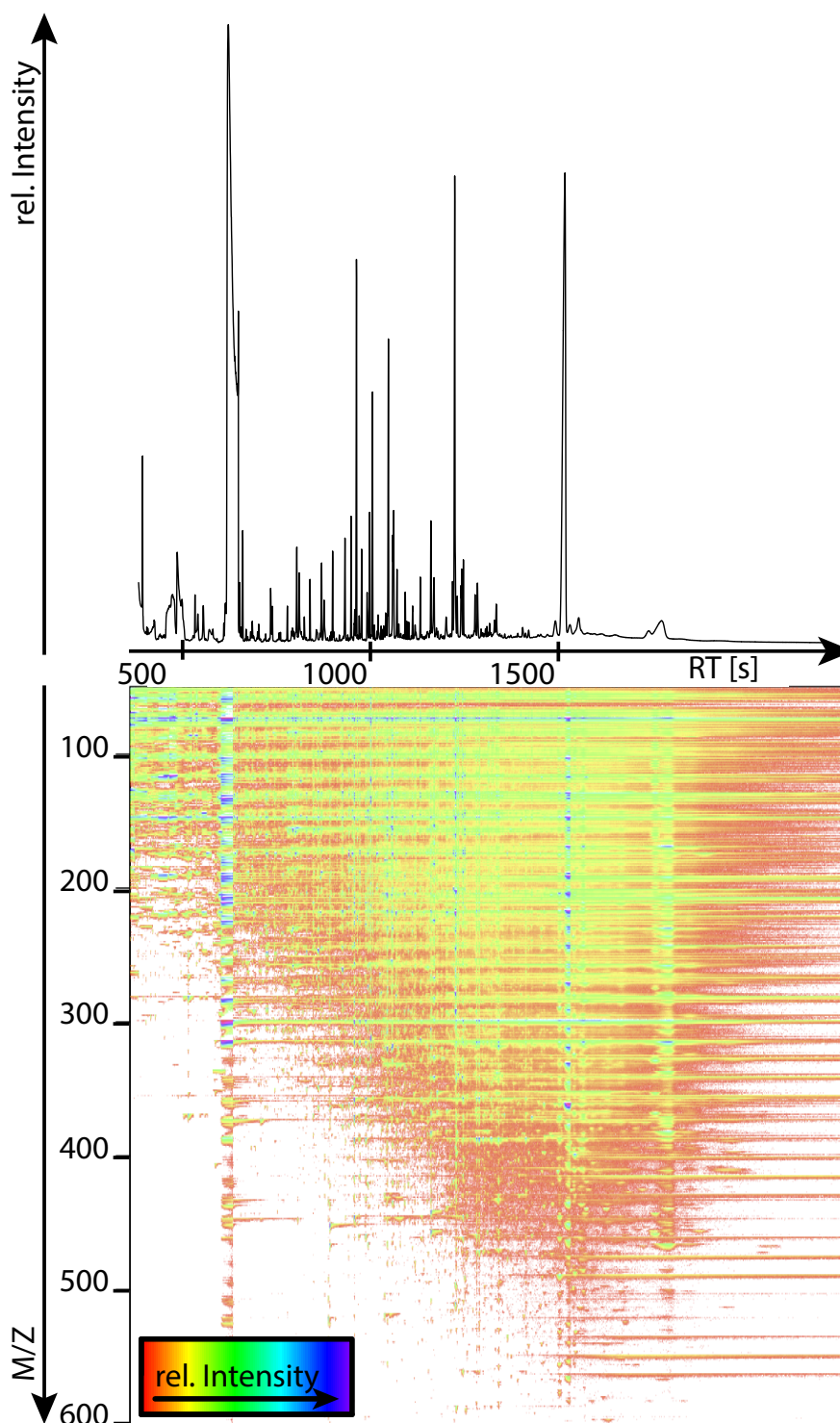


Figure A3.4.: **GC-MS spectrum:**

2D heat map of a GC-MS acquisition of a *C. elegans* extract with the corresponding TIC on top. Underneath intense TIC peaks, their fragmentation pattern is visible in the heat map.

algorithms have been developed. Even though some of these are very complex, involving probability calculations and using meta information [404], the simple dot product matching algorithm, performs unchallenged in routine matching applications [540]. The dot product matching algorithm simply calculates the dot product between two unit norm scaled intensity vectors of an unknown ($V_{Unknown}$) and a library ($V_{Library}$) spectrum, according to equation A25.

$$\frac{(\sum |V_{Library} V_{Unknown}|)^2}{\sum V_{Library}^2 \sum V_{Unknown}^2} \quad (A25)$$

In case both vectors match perfectly, *i.e.* the vector is identical, the dot product becomes one. If both vectors contain no intensity at a single, common m/z value, the dot product will be zero. Partial matches will therefore yield scores between 0 and 1, depending on the two vector's similarity. When performing a database query for a fragmentation pattern, the software would compute the dot product between the queried vector and all database fragmentation patterns and rank them according to their similarity. If the match score exceeds a prior set threshold (and the retention time index is also within a set confidence interval), the fragmentation pattern may be assigned as the best matching compound. As the dot product is a simple arithmetic procedure, the algorithm is very fast and allows large databases to be queried quickly.

The most widespread used matching software was also developed by NIST: AMIDS, the Automated Mass Spectral Deconvolution and Identification System, developed by Stein and coworkers [539] is actually much more than a simple fragmentation pattern matching software. It allows to find the fragmentation patterns in the first place, *i.e.* it determines which fragments belong together and are likely to originate from a single compound. Those extracted m/z clusters are called components, which then can be matched to a database. Not only is it highly sensitive, but AMDIS is also able to deconvolve overlapping fragmentation patterns, which originate from compounds, that were not completely separated by the gas chromatograph. AMDIS will process a sample set on a file-to-file basis, outputting the detected components, their intensity and their database annotations, in case a compound could be matched. However, when used for metabolomics experiments with multiple sample sets, the AMDIS workflow has two shortcomings, compared to the xcms software package used for LC-MS processing (A3.1.3.1). It does not align the components across multiple files. Neither does it allow to re-integrate the area, in cases where a component was not detected in a particular sample.

For this reason, post processing workflows are required. The standard (post processing) workflow used for processing the GC-MS data sets in this thesis is presented in the next section.

A3.1.4.2. GC-MS processing workflow

First, the acquired data files are converted into the open source netCDF (Network Common Data Form) format, using the vendors software. Prior to its use, AMDIS requires adjustments to its parameters in order to deliver the best possible results. In particular, the average peak width (scans per component), minimum intensity threshold and the threshold for the fragmentation pattern and retention index matching require to be optimised. This is best done by using a QC sample as a representative example and varying the parameters until a satisfactory performance is yielded. Satisfactory would be identifying and matching a high number of peaks, which can be assumed to be matched correctly. After those settings are found, AMDIS is executed on all files. To align the annotated components across multiple samples and re-integrate peaks that were missed by AMDIS, a software complement, in-house developed by Volker Behrendts *et al.*[38] was used. The Matlab based software called GAVIN, steps through each metabolite individually, visualising the AMDIS integration boundaries and allowing to adjust them manually. Furthermore, it allows checking of the assigned mass spectra to the spectral database in order to exclude potential miss-matches. It is also possible to select the quantification ion - the ion of the component (fragmentation pattern) which is used as the representative intensity for this component. In case of overlapping components, the fragmentation ion must be chosen to be unique to the desired component. GAVIN outputs an intensity metabolite $\times$ sample matrix, which can be used for subsequent data analysis, as described in the next section.

Even though this workflow has proven to work well, it has a major drawback. Only components, that were identified as a metabolite (by the database matching), are considered. Components that are not contained in the database or can not be matched with sufficient confidence are discarded. Numerous other GC-MS data processing workflows have been reported [144, 172, 570, 63, 548, 83, 95], that (partially) address this problem.

Chapter III.4 will develop a GC-MS workflow, which works irrespective of database matching and is a TIC alignment WYSIWYG (what you see is what you get) approach to GC-MS metabolomics.

A3.2. Data analysis and statistics

After peaks have been extracted from the raw data, they may be analysed using classical univariate statistics or multivariate pattern recognition techniques, described in section A3.2.1 and A3.2.3, respectively. Section A3.2.2 gives a short overview over data preprocessing methods, such as normalisation and scaling, which are of particular importance, when multivariate models are used.

A3.2.1. Univariate statistics

This thesis will not only use statistics to analyse the results and test upcoming hypothesis, but also the methods, with computational algorithms that are used or developed, in particular, rely on key statistical concepts [70]. Therefore important statistical concepts will be revised briefly in this section.

After a short revision of important parameters of descriptive statistics, inference statistics, with confidence intervals and statistical testing will be discussed.

A3.2.1.1. Distributions and descriptive statistics

Any event in the universe has a certain likelihood for it to occur - its probability. It may not ever occur, and thus its probability will be infinitely close to 0 (0%) or inevitable, with a probability of 1 (100%). Probabilities are sampled from their underlying probability distributions. Even though a probability distribution can have any possible shape, there are a few model distributions, that most processes can be assigned to. Those model distributions, either discrete or continuous, can be described by mathematical functions. Their probability density functions (PDF). Import distributions include the Binomial, Boltzmann, Chi-squared and Poisson distribution. The distribution mostly used in this thesis and maybe science in general, is the normal distribution, as shown in figure A3.5.

If however, the type distribution cannot be determined to make inferences about its parameters, nonparametric methods, such as those that rely on ranks, rather than a continuous scale, may be used [188]. A very important non-parametric way of describing a distribution

is the median or any other percentile. If ranked, the median is the numerical value at half of the data sample. The median is the 50th percentile. The 95th percentile, would be the numerical value of the 95th of 100 ranked values.

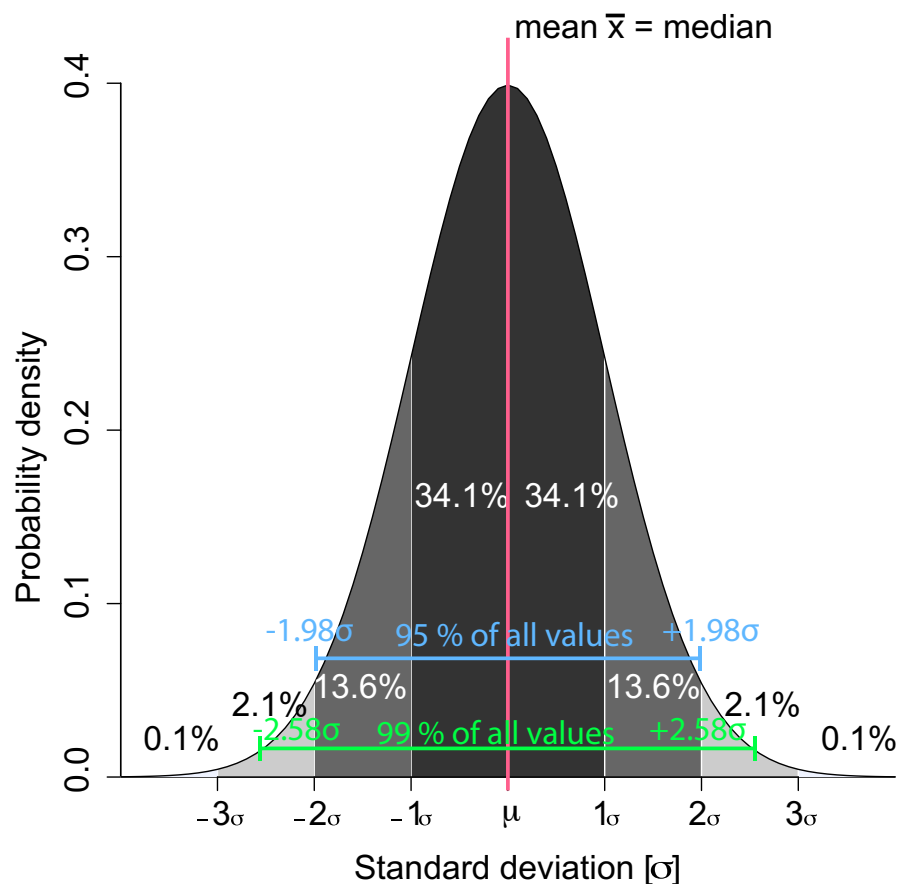


Figure A3.5.: **The normal distribution:**

The probability density function of a standard normal distribution, *i.e.* a normal distribution with a mean μ of zero and a standard deviation σ of one. Also, in a non-skewed (normal) distribution, the mean equals the median and has the highest density. The often used 95 % and 99 % confidence intervals are also shown.

According to its PDF, given by equation A26, the normal distribution, can be characterised by additional parameters, also illustrated in figure A3.5. Its mean μ and variance σ^2 , which often is expressed as the standard deviation $\sqrt{\sigma^2} = \sigma$. Most parametric statistical tests build on the normal distribution and therefore these two parameters.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (\text{A26})$$

For a set of n values $x_1, x_2, \dots, x_i$, the sample mean $\bar{x}$ can be calculated according to equation A27.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (\text{A27})$$

For the same set, the sample standard deviation (SD) can be computed according to equation A28.

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (\text{A28})$$

In non skewed normal distributions, the mean equals the median and maximum of the PDF, specifying the most likely value from this distribution. The standard deviation describes the amount of variation inside a distribution, in other terms the dispersion from the mean. The larger, the further away individual samples can be found away from the mean. Thus, it can be used to build confidence intervals, as will be explained in the next section.

Another measure that will be used frequently throughout the thesis, to describe the reproducibility of an analytical measurement is the relative standard deviation (RSD), also referred to as coefficient of variation (CV). According to equation A29, it is the ratio between the standard deviation SD and the mean $\bar{x}$.

$$CV = \frac{SD}{\bar{x}} \quad (\text{A29})$$

A way to express the strength of the correlation between two vectors, is to calculate their covariance, according to equation A30.

$$cov(X, Y) = \sum_{i=1}^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N} \quad (\text{A30})$$

If the covariance is normalised by the standard deviations of both vectors, the Pearson's correlation coefficient R is yielded (equation A31).

$$R = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \quad (\text{A31})$$

$|R|$ describes the magnitude of the two correlating vectors, with 1 being co-linear and 0 being orthogonal. If R is positive and negative, the correlation is direct and indirect, respectively.

A3.2.1.2. Statistical inference

Statistical inference encompasses all the methods and concepts, that can be used to draw conclusions from data that is influenced by random variation and therefore an integral part of science. In order to choose the right tool of statistical inference, it is important to know the underlying structure of the data, such as from which probability distribution it is derived or whether it has been treated as non-parametric. Statistical inference concepts divide into Bayesian inference and frequentist inference. Only the latter will be discussed in the scope of this section as it is the one immediately relevant. Furthermore, unless noted otherwise, a normal distribution is assumed for all further considerations.

When during an experiment a reading is performed, a sample (values) from the underlying population distribution is taken, and used to build up a sampling distribution. Ideally the population and sampling distribution are the same. However, in reality this largely depends on how many random samples (n) are taken. The more - the more the sampling distribution's mean and standard deviation will converge to the ones of the population. Figure A3.6 illustrates this.

If sampling from the population's distribution is repeated, each time, the sampling distribution's mean and standard deviation will be different, varying around the true values. The extent to which repeated samples vary around the true value (*i.e.* the standard deviation), can be estimated using the standard error of the mean (SEM). According to equation A32, the SEM depends on the population distribution's standard deviation s and the number of samples n taken in during the sampling procedure. Those concepts show the importance

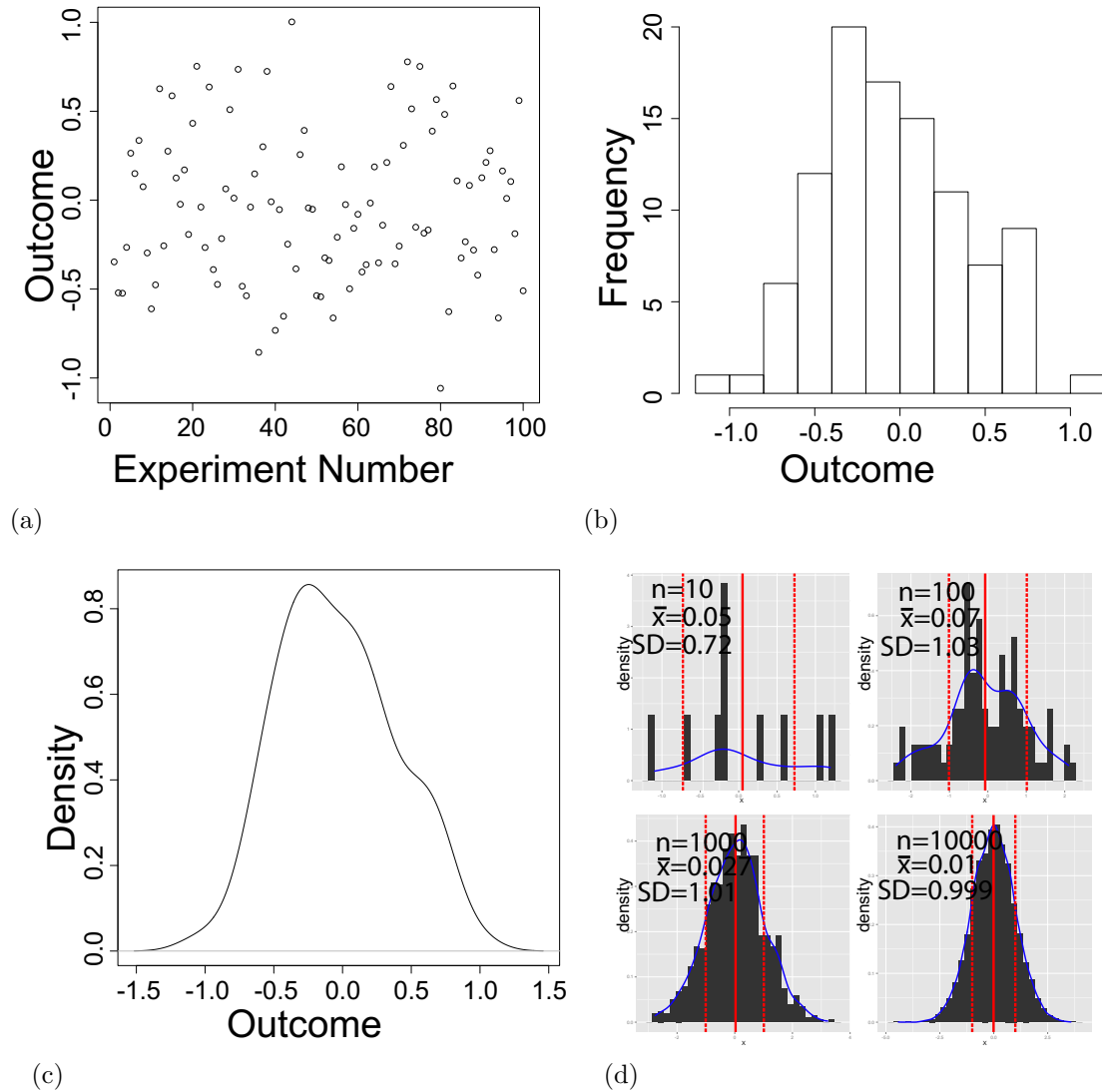


Figure A3.6.: **Probability and its distribution:**

During an experiment 100 samples were taken from a normally distributed population distribution (figure A3.6a). The results can be summarised in a histogram - bucketing the outcomes and the frequency of those outcomes (figure A3.6b). A histogram is a discrete, empirical probability distribution. It is also possible to fit a continuous probability density function to the sampled values (figure A3.6c). Figure A3.6c shows the histograms and fitted pdf in for four different sample sizes. It is obvious that the more samples, the closer the empirical pdf resembles the pdf of a normal distribution, with sampling distribution's parameters (SD as dashed line; $\bar{x}$ as solid line) approaching the population distribution's parameters.

of having replicates, *i.e.* a certain sampling size, in any scientific experiment.

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (\text{A32})$$

In a normal distribution, the SEM can be used to estimate confidence intervals. If multiplied with 1.96, then added and subtracted from the sample mean, the upper and lower 95% confidence interval are obtained, respectively. Confidence intervals are an important aspect of inference statistics, as they give an estimated range, where the unknown population distribution's parameters are expected. For example, when a sampling procedure, of several samples each, is repeated 100 times, 95 of the 100 different sampling distribution's parameters should lie within a prior defined 95 % confidence interval. Thus, a confidence interval allows to estimate whether it is likely to have obtained a realistic estimate of the population parameters during a sampling event. This is the basis for statistical hypothesis testing [176].

In statistical hypothesis testing a decision is made, whether a result could have occurred by chance alone, *i.e.* whether it is or is not from a certain population distribution, defined by a confidence interval. The general procedure is as follows:

A certain hypothesis, for example, whether a biomarker is associated with a disease is put forward. The null hypothesis H_0 would state that there is no association, whereas the alternative hypothesis H_1 claims there is. A significance test aims to test whether H_0 can be rejected with a certain confidence, in favour of H_1 . It is essential to know the structure of the data, such as underlying type of distribution or statistical independence, in order to choose the correct test, when making the inference. In this thesis, only two major groups of tests will be used. T-tests and F-tests (ANOVA). (Exact details about the testing procedures, such as used test and significance levels will be given in the method section of the corresponding results chapter.) T-tests are a group of significance tests that use the mean, whereas F-tests use the variance to judge whether H_0 may be rejected. Both work by calculating a T-statistic and F-statistic, that are compared with a T-distribution and F-distribution, respectively, using lookup tables to yield a p-value.

P-values are the probability of obtaining a test statistic at least as extreme as the calculated one. The lower, *i.e.* the closer to the tail of a distribution, the less likely. Geometrically,

a p-value is the integral of the probability density functions tail, starting at the observed data point and ending at infinity. If a p-value is smaller than the prior agreed significance level α , H_0 is rejected in favour to H_1 . A common significance level is 5 %, which would correspond to a 2σ event in a normal distribution (compare figure A3.5).

According to contingency table A3.2 four scenarios are possible, but only two of them desired; accepting H_0 when H_0 is actually true, or rejecting it when it is actually false.

	H_0 is true	H_1 is true
Accept H_0	Right decision True positive	Wrong decision (Type 2 error) False negative
Accept H_1	Wrong decision (Type 1 error) False positive	Right decision True negative

Table A3.2.: *Contingency table of statistical hypothesis testing outcomes*

A type 1 error, also referred to as false positive, is when H_0 is incorrectly rejected. A relationship is falsely assumed. In the prior example, the biomarker would be associated with the disease, albeit in reality it isn't. In fact, the probability of committing a type 1 error, equals the significance level α . Type 2 errors, false negatives, accept H_0 , even though H_1 is true. The test failed to reveal an actually correct relationship. For example, the biomarker would not be associated with the disease, even though it is. The probability β of committing a type 2 error is unknown. However, $1 - \beta$ is the power of a hypothesis test, *i.e.* its probability of not committing a type 2 error.

In hypothesis testing, it is important to find the right trade of between these two errors. Thus, choosing an appropriate significance level is important in order to neither miss, nor state incorrect relationships.

In metabolomics, more than one compound are measured and evaluated for their probability of being a (bio)marker, associated with a certain condition, such as disease or manipulation. This would require the calculation of multiple statistical hypothesis tests. But if one were to perform 100 tests with a significance level of $\alpha = 5\%$, on average, one will commit five type 1 errors - without knowing which in which instances these errors occurred. To avoid this, one can acknowledge this problem and adjust the p-values by multiple testing procedures [40], such as Bonferoni [3] or Benjamini-Hochberg correction

[41]. Bonferroni is the more conservative method and tries to control the family-wise error rate (FWER), *i.e.* the probability of committing one or more type 1 errors. This can be done by dividing the initial significance level, *e.g.* $\alpha = 5\%$ by the number of performed tests n . In case three performed tests, $5\%/n = 5\%/3 = 1.7\%$ would be the new α . Benjamini-Hochberg correction controls the false discovery rate (FDR), *i.e.* the expected proportion of type 1 errors among all significance tests. It is less conservative as it tolerates more false positives, but reduces the number of false negatives (increased power). To perform it, p-values are sorted from the largest to the smallest. The largest p-value remains the same, but the second largest is adjusted by multiplication with $n/(n-1)$, the second largest with $n/(n-2)$, and so on.

Another strategy to minimise type 1 errors in multiple testing scenarios would be by minimising the number of necessary comparisons in the first place. For example, by using multivariate data exploration tools or multiple group statistical testing procedures, such as ANOVA, the total number of useful tests may be constrained to the most promising candidates only, rather than testing every possible combination.

ANOVA is a collection of linear statistical models, that compare the total variance to the one of individual groups to derive a F-statistic and p-value. If the p-value is significant one of the groups is significantly different. The different group can be found using post-hoc testing procedures.

Multivariate models, such as principal component analysis, will be discussed in section A3.2.3. However, multivariate methods often require the prior use of normalisation, centering and scaling procedures discussed in the next section.

A3.2.2. Data preprocessing: Normalisation, centering and scaling

Irrespective of subsequent analysis, data should be normalised [599]. Normalisation is usually the first step of data pretreatment. It aims at reducing concentration effects that arise from a bias in sampling, such as different concentration in extracted biomass or dilution effects (cell media, urine) [115]. Ideally, after normalisation, only differential metabolites would differentiate two samples, rather than one's overall higher concentration (intensity). Three different approaches for normalisation are typically used. Scaling to external standards, scaling to internal standards and empirical approaches. An example for scaling to

an external standard would be using another, from the profiling method independent assay, such as a spectroscopic Bradford protein determination to compensate for differences in sampled biomass. As internal standard, a metabolite that is known to be contained in a constant concentration inside a given matrix, could be used - analogously to the "house-keeping genes" used for the normalisation of microarray expression data. An example for an internal standard scaling factor is creatinine in urine samples, as its excretion can be assumed as constant [281]. Alternatively, if normalisation is only desired to compensate for distortions occurred during the sample preparation, such as incomplete recovery during extraction or pipetting inaccuracies, spiked in (stable isotope labelled) standards could be considered. Two main empirical approaches have been used; total integral normalisation and probabilistic quotient normalisation (PQN) [136]. Whereas the first has come under scrutiny and should be avoided, particularly for MS detected datasets, the second one will be used in this project. PQN works under the assumption that the median fold change of all variables equals zero. Unless in the rare hypothetical case that the vast majority of signals displays a fold change in one particular direction (most metabolites are either up- or downregulated), it is very robust. Therefore it has become the default normalisation method for metabolic profiling studies.

Mean centering and scaling are the two next pretreatment steps, which are very common, particularly when multivariate analyses are used. Their equations are given in table A3.3. Mean centering sets the mean of each variable across all samples to zero, by subtracting the mean. Therefore all changes are centred around zero, rather than the mean of a metabolite. Thus if a variable wouldn't change across any of the samples, it is set to zero and therefore "removed" from the analysis. Also, the offset between high and low intense features is removed. Thus, only relevant concentration changes are left for analysis. Figure A3.7 illustrates its effect on variables.

Scaling divides all variables by a number - the scaling factor. Its goal is to reduce large differences in fold changes across the different variables. Otherwise, inside a multivariate model, a very high fold change of one variable could outweigh several smaller, but equally important fold changes of other variables. However, this may also result in the danger of allowing small, random, noise derived variables to have a higher weight in the analysis and lead to misinterpretation. Therefore it is particularly important to follow up with (unscaled) univariate statistical testing, when scaling methods were used.

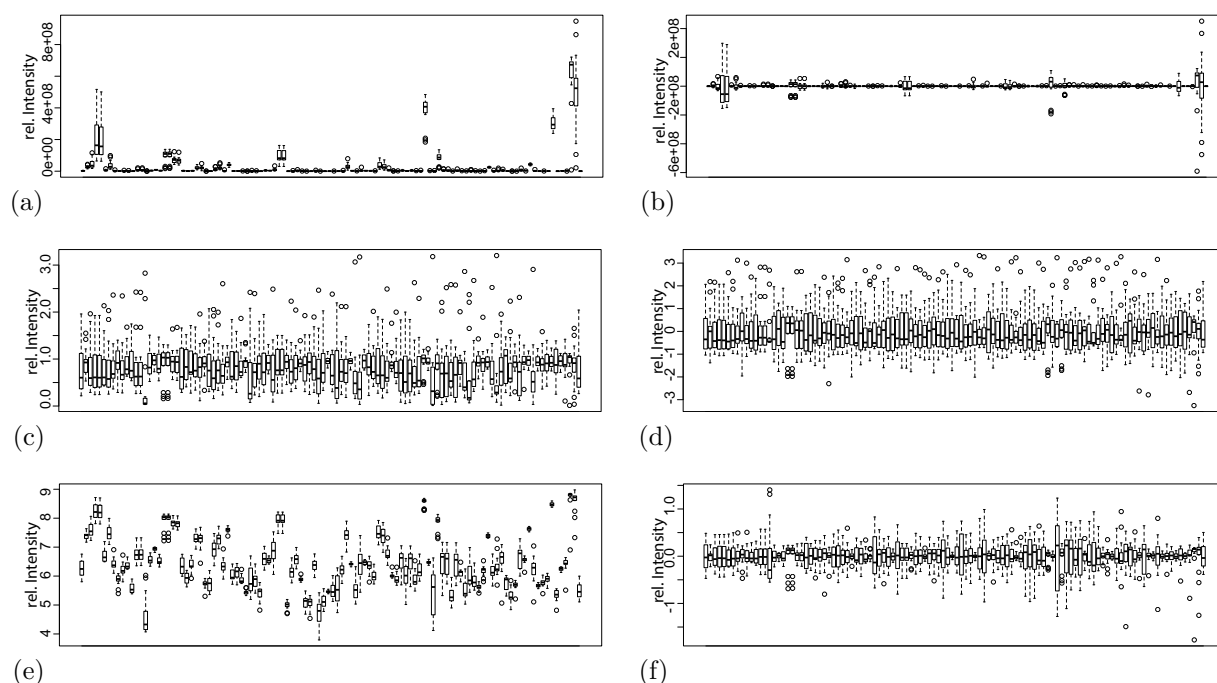


Figure A3.7.: **Effect of mean centering and scaling on data:**

Boxplots of metabolites of a typical metabolomics experiment are used to visualise the effect of mean centring, unit variance and log scaling. The right column is the mean centered version of the left one. Figure A3.7a shows the raw data, figure A3.7c the unit variance scaled and figure A3.7e the log scaled (base 10).

The two most common scaling methods are unit variance scaling and pareto scaling. Unit variance scaling uses mean centred data and divides it by its standard deviation. This results in a standard deviation of one for all data, and therefore gives all variables equal importance (figure A3.7c). Also, it allows data to be analysed as correlations, rather than covariance. Pareto scaling is slightly more conservative. It divides by the square root of the standard deviation. Thus large fold changes are scaled more, relative to smaller ones. Therefore, compared to unit variance scaling the data stays closer to its original structure.

Albeit often referred to as a scaling method ("log scaling"), a logarithmic transform is technically no scaling method, as it does not divide by a scaling factor. However, it usually is used with the same intention as the methods mentioned above (figure A3.7e). Care must be taken when the standard deviations are high as log is a logarithmic transform thereby

Operation	Type	Formula	Unit
Centering	Centering	$\hat{x}_{ij} = x_{ij} - \bar{x}_i$	original
Unit variance	Scaling	$\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	dimensionless
Pareto	Scaling	$\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	original
Log	Transform	$\hat{x}_{ij} = \log_{10} x_{ij}$	log original
started Log	Transform	$\hat{x}_{ij} = \log_{10}(x_{ij} + \min(abs(x)))$	log original
generalised log	Transform	$\hat{x}_{ij} = \ln(x_{ij} + \sqrt{x_{ij}^2 + \lambda})$	dimensionless

Table A3.3.: **Data preprocessing methods:**

Formulae for the various data preprocessing methods explained in the text [587]. i and j are the sample and variable index, respectively. $\hat{x}$ denotes the data after transformation. The transform parameter λ is discussed in the text.

giving different scaling weights to different numbers. Log transformation is also useful to reduce heteroscedasticity and transform multiplicative interactions to additive relations. A major disadvantage of log transform is that it is not defined for negative numbers. Therefore mean centering has to be performed after the transformation. Furthermore, if there are any negative values in the dataset, it is not applicable. However, for this case two closely related alternatives have been formulated. The started log transform [500] and the generalised log transform [146, 145, 464]. For the started log transform a small offset, usually the absolute value of the smallest (negative) value, is added. Then a regular log transform can be performed. As the generalised log transform yields similar results, but requires a complex calibration of a sample set specific parameter (λ) [464], the started log transform is used in this thesis.

A3.2.3. Multivariate statistics and pattern recognition

Multivariate techniques are useful, when a system has more than one outcome variable. For example, if a cohort of patients should be compared with regards to their blood glucose levels, univariate methods are appropriate. If, however, the patients should be compared, based on the measurement of their 22 amino acids, multivariate methods should be used, instead of performing 22 univariate tests; one for each amino acid. This is because 22 T-tests, for example, would drastically increase the likelihood for a type one error or otherwise require very stringent multiple testing correction. A multivariate technique would already hint at the relevant variable, which could be followed up with univariate post-hoc tests.

Particularly in metabolomics experiments, where the number of variables may be as high as a few ten thousand or more, this is an important consideration. The most classic multivariate method is a generalised version of the univariate ANOVA, the multivariate analysis of variance (MANOVA) [533]. However, like for ANOVA, it performs inference testing based on probability distributions and therefore has certain prerequisites to its input data quality. In metabolomics datasets, those prerequisites are not always met. Furthermore, a MANOVA is very powerful for analysing factorial designs, but less suited for the initial exploratory analysis of a large dataset [285]. Therefore the metabolomics community usually defaults to multivariate pattern recognition techniques for the initial analysis. [521, 275]

Pattern recognition techniques are used to classify data and reduce the number of its dimensions. The number of dimensions can be reduced by expressing the data through latent variables. Latent variables are weighted linear combinations of the original variables. In the simplest case, if a variable (dimension) does not discriminate between any data points, it may as well be discarded. Particularly in metabolomics experiments, data often displays high redundancy and multicollinearity, such as one compound described by multiple resonances in NMR or one compound giving rise to multiple isotope and adduct peaks in MS. This redundant information can be summarised by the use of latent variables and the number of variables (dimensionality) reduced. [579]

Pattern recognition techniques divide into two categories - supervised and unsupervised methods. Supervised methods, have *a priori* information about the data's metadata, such as the sample's group membership. Unsupervised methods do not have any information about the system whatsoever. They entirely rely on the data's internal structure, such as its variation, distance or correlation, when analysing the relationship between individual observations (samples). The most common supervised method in metabolomics are derivatives of partial least squares regression (PLS). Used unsupervised methods are hierarchical cluster analysis (HCA) and the metabolomics workhorse - principal component analysis (PCA). [384]

A3.2.3.1. Principal component analysis (PCA)

Principal component analysis (PCA) [471] is a projection technique, that allows to visualise

the maximum variance, encountered in the data.

First, the data matrix X requires to be mean centred [416], *i.e.* its variable means must be zero, as described in section A3.2.2.

The m by n data matrix X , has m number of observations (samples) and n number of variables (metabolites, peaks or spectral bins). PCA will transform this data into a new coordinate system. The principal component space. Thereby, the original matrix X is decomposed into a scores matrix T and a component matrix P , according to equation A33. E is the error term, summarising all residuals. The scores matrix T contains the orthogonal latent variables (principal components), with most of the variance described by principal component one ($t_1p_1^T$), then two ($t_2p_2^T$), and so on. P , the loadings matrix, describes the influence (weight), of the original X -variables inside the principal component space.

$$X = TP^T + E = t_1p_1^T + t_2p_2^T + \dots + E \quad (\text{A33})$$

The transformation into the principal component space can be best explained in a geometric representation. Each observation of the original matrix X , can be considered as a datapoint inside a n -dimensional hyperspace (n -dimensional coordinate system), spanned by the variables of X . This as illustrated by figure A3.8a for a $n = 3$ scenario.

PCA transforms all these data points into a new coordinate system, spanned by latent variables, the principal components. The first principal component t_1 lies in the direction of the largest variation. t_2 explains as much possible variation, that has not already been accounted for by t_1 , *i.e.* it is orthogonal to t_1 . This continues until all variation of the system has been explained by principal components. Figure A3.8 illustrates this procedure.

Compared to the interpretation of a high dimensional data matrix, interpretation of the principal components, particularly if plotted, is straight forward. As the principal components were defined along the paths of largest variation, when plotted against each other (in the "scores plot"), the principal components display the largest variation. In the simplest case, the largest variation within a dataset is between two distinguishable groups, such as diseased and healthy, or wild type and mutant. In such a case, the groups are expected to separate along one or a combination of the principal components. Principal

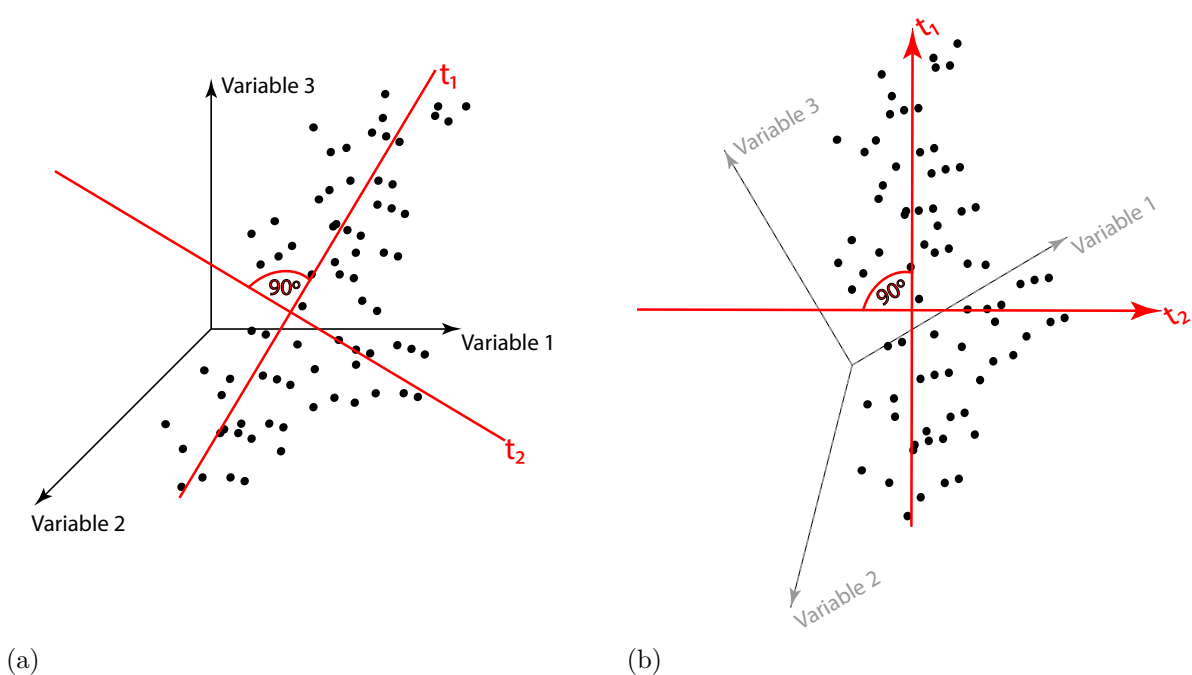


Figure A3.8.: **Principal component analysis (PCA):**

Each observation is a data point inside a n-dimensional hyperspace. In this case a data matrix with $n=3$ different variables (such as metabolites). The principal components are spanned in the direction of the highest variation and are orthogonal to each other. Transformed data points plotted inside the principal component space are called scores plot (figure A3.8b).

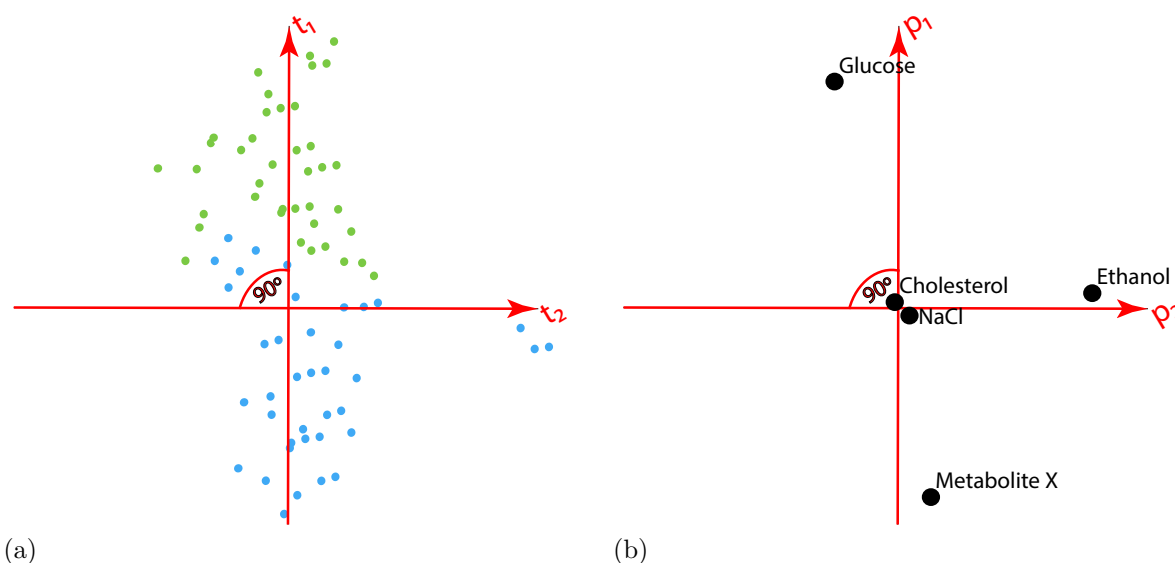


Figure A3.9.: **PCA scores and loadings plot interpretation:**

This is a fictional example explaining the interpretation of the scores and loadings plot, yielded by a principal component analysis. Consider a blood test being performed on a cohort of healthy and a cohort of diabetic patients. Five compounds are measured: Glucose, cholesterol, sodium chloride, a yet unidentified metabolite X, and the ethanol levels, to check the patient's compliance. After principal component analysis, the scores plot, displayed in figure A3.9a, shows that the largest variation in the dataset lies between the diabetic (green) and healthy (blue) individuals. Those clearly separate in direction of t_1 , with a group of three healthy, as outliers on t_2 . Consulting the loadings plot, in figure A3.9b, the variables driving the separation can be identified. As expected, the diabetic patients are characterised by higher glucose levels. Cholesterol and sodium chloride seem to be irrelevant, when discriminating between diabetic and healthy. Metabolite X however, is of higher concentration in the healthy population and it would be interesting to annotate it. Also, the three outliers show bad compliance to the study design, as they seem to have consumed ethanol and may therefore be excluded from subsequent statistical analyses. Next to elucidating biomarker X, univariate methods, for example a t-test, should be used to assess, whether it is significantly associated with the healthy individuals.

In this example, the five measurement dimensions were reduced to three relevant ones, and if outliers are removed, to two. This illustrates the power of PCA for high dimensional metabolomics data. There is also a general spread in t_2 , even though, despite the three outliers, the other study participants may be sober. PC are latent variables. Glucose and metabolite X mainly separate on t_1 , but also have a small impact on t_2 , thereby giving all observation a spread and partial diagonal character, rather than making them a singular line on p_1 .

components are latent variables, *i.e.* composed out of the weighted original variables. To find out, which original variable is discriminating between the two groups (visible in the scores plot), these weights, contained in the loadings matrix can be plotted. The resulting "loadings plot" contains all original variables, with their positions arranged according to the weight they have for the principal components in the scores plot. Thus, if a variable in the loadings plot is spatially aligned with a separated group in the scores plot, this variable causes the separation. If a variable is close to the centre of the loadings plot, its impact on the plotted principal component is negligible. In this manner trends, trajectories and (dis)similarity between the sample groups can be explored. Figure A3.9 illustrates this by giving a fictional example.

It is important to note, that when explaining the loadings from a t_1 versus t_2 scores plot, the associated p_1 versus p_2 loadings plot must be used. Similarly, the question, how many principal components should be interpreted, emerges. In theory, there is no limit, except, that at most, there are as many principal components as initial variables. However, the information content (with regards to variation) drops as with each PC. The percentual amount of the system's variation explained by a single PC is often given as its R^2 value. As the amount of explained variation decreases with higher principal components, the R^2 does so as well ($R_1^2 > R_2^2 > \dots > R_n^2$). In most cases, most of a system's variation can be explained within the first few PC ($R_{1-3}^2 = R_1^2 + R_2^2 + R_3^2$). In this thesis, only the first two PC will be used.

$$R^2 = 1 - \frac{\sum E^2}{\sum X^2} \quad (\text{A34})$$

R^2 is calculated according to equation A34.

Finally, it is important to keep in mind, that ordinary principal components or R^2 values are not associated with significance values [89], as they only are projection techniques. In scenarios like the example from figure A3.9, they should always be followed up using inference statistics.

A3.2.3.2. Partial least squares regression (PLS)

Partial least squares regression (PLS) is a supervised modelling technique. It expresses the

common structure between the X and Y matrix, by maximises their covariance and can therefore be used for prediction. PLS decomposes those matrices according to equation A35 and A36, with T and U as score matrices, P and C as loadings matrices. E and F contain the unmodelled residuals. [579]

$$X = TP^T + E \quad (\text{A35})$$

$$Y = UC^T + F \quad (\text{A36})$$

In this thesis, Y is not a matrix, but a vector, such as life expectancy for the observations in the data matrix X . The scores plots (usually only U is used) can be interpreted similar to the one's of PCA. However, it must be kept in mind that it is not variation, but covariance that the separation is based on. Also, as it has *a priori* information, *i.e.* is a supervised method, it is important to validate the model and make sure it is actually predictive and not over fitted [622]. To this end, several methods for validation are available (reviewed by Szymanska *et al.*[156]). The most commonly and also the approach used in this thesis, is Q^2 cross validation. Q^2 is the cross-validated version of R^2 , and therefore tells how much of the covariance can actually be predicted by the model. If a model has a high R^2 and low Q^2 value, it fits the data, but is not predictive, *i.e.* over fitted. Q^2 can be derived via leave-one out methods. One sample (group) is excluded from building the model, but used during prediction of $\hat{Y}_{cv}$. After every sample (group) has been left out once, Q^2 can be calculated according to equation A37.

$$Q_Y^2 = 1 - \frac{\sum(\hat{Y}_{cv} - Y)^2}{\sum Y^2} \quad (\text{A37})$$

A3.2.3.3. Partial least squares - discriminant analysis (PLS-DA)

PLS regression can also be used for discriminating between different sample groups, *i.e.* force the separation between those in a supervised manner. In this case, Y is used as a dummy variable, corresponding to the sample classes, for example $Y_i = 0$ for control, $Y_i = 1$ for wild type. Again, cross-validation determines if the model is actually useful. [579]

Protocols

B1. Wormculture

All protocols were executed, according to the Wormbook - a peer reviewed online resource of the *C. elegans* community [544]. When necessary, deviations from the standard protocols were made. Chemicals and consumables for the worm culture were obtained from VWR (Lutterworth, UK). Exceptions are stated. All worms were grown at 20°C inside a cooled incubator.

B1.1. Growth medium

Worms are cultured on nematode growth medium (NGM) plates. Those are prepared by dissolving 3 g NaCl, 17 g Agar and 2.5 g bacterial peptone in 0.975 L of water, purified by a millipore system (Millipore, Watford, UK). After autoclaving and letting cool to $\approx 60^{\circ}\text{C}$ the following autoclaved solutions are added: 1ml 1M CaCl_2 , 1ml 1M MgSO_4 and 25 ml 1M phosphate buffer (108.3g KH_2PO_4 , 35.6g K_2HPO_4 , H_2O to 1 litre) adjusted to pH 6. Finally 1 ml of a 5 mg/ml cholesterol in ethanol solution subject to sterile filtering through a $0.22\mu\text{m}$ PTFE filter are added, prior to pouring into 90 mm petri dishes ($\approx 20\text{ml}$ per plate).

Plates are left at room temperature for two days to allow inspection for possible contamination. Then, they are either stored at 4°C for later use (up to a maximum of 4 weeks) or seeded with 200 μl of an overnight culture of *E. coli* strain OP50, grown in 25 g/l luria broth medium in H_2O . Seeding is done using a l-shaped spreader to create a lawn of uniform thickness, staying clear of the edges of the petri dish. Plates are incubated overnight at 36°C to allow for a bacterial lawn to grow.

B1.2. Age synchronisation

In some experiments, a mixed-age population was used to maximise metabolic diversity. However, to obtain a clear snapshot of a metabolic phenotype, age synchronisation is required. This can be achieved by a wormbook[544] procedure called "bleaching". Gravid worms are subject to an alkaline bleach solution to "dissolve" them and release their eggs which hatch at the same time $\pm$ 4 hours.

The worms of an agar plate, containing several gravid hermaphrodites, are washed into a 15 ml conical tube, using 2 x 3 ml aliquots of sterile M9 buffer (3 g KH_2PO_4 , 6 g Na_2HPO_4 , 5 g NaCl, 1 ml 1 M $MgCl_2$, H_2O to 1 litre). 1 ml of 10 M NaOH and 3 ml of 10 % NaOCl solution are added and topped up to 10 ml with dd H_2O . After shaking for 5 minutes the tube is centrifuged for 2 minutes at 2000 g. The supernatant discarded, and the pellet washed twice with 10 ml of M9. Before the pellet is transferred to the edge of a newly seeded NGM plate, the pH of the last washing supernatant is checked to ensure it is at pH 6.

B2. Default sampling and extraction protocol

Ahead of any profiling experiments, an extraction method evaluation, outlined in chapter II., was performed. The consensus from these experiments, led to a standard sampling and extraction protocol used for all subsequent analyses.

To prevent degradation, all samples were kept on dry ice as much as possible throughout extraction. Ice cold only solvents were used. All solvent and materials used for the extraction were obtained in highest available purity.

Based on the results from the extraction method comparison, the following standard protocol for all future extractions was derived:

For harvesting worms of a single agar plate (corresponding to one biological replicate), those were washed into a 15 ml conical tube, using 2 x 3 ml aliquots of sterile M9 buffer. After 5 minutes, all worms will have settled as a pellet on the bottom of the tube. At this stage a washing step could be performed by letting the pellet settle inside a fresh tube filled with 10 ml of M9 buffer. The necessity of a washing step is discussed in II.5.1. However,

in most experiments no washing steps were performed, but deviations will be mentioned. The (washed) pellet of 300 μ l was transferred into a 2 ml bead-beater tube. This was done using a glass pipette, to minimise the loss of biomass, due to an observed static adhesion of the worms to plastic pipette tips. The pellet was snap frozen immediately and stored at -80°C until extraction.

On dry ice, the bead-beater tube was filled with 200 μ l of 0.1 mm glass beads and 1.2 ml of cold methanol were added. The sample was immediately extracted using a Precellys 24 bead-beater (Bertin Technologies, Montigny-le-Bretonneux, France), operated for two 20 second cycles at 6.5×10^3 rpm, with 20 seconds pause inbetween. Samples were centrifuged (5 min, at 16×10^3 g, 4°C), the supernatant transferred into a 15 ml centrifuge tube and the pellet re-extracted with 1.5 ml of 80% methanol. The obtained supernatant was pooled with the first and the total volume topped up to 3 ml - where necessary. This volume was aliquoted into three 1 ml technical replicates, which were then dried down in a vacuum concentrator at room temperature and store at -80°C until analysis. Aliquots designated for GC-MS analysis were dried directly inside silanised GC-MS glass vials (Agilent), rather than 1.5 ml sample tubes.

B3. NMR spectroscopy

B3.1. Preparation

Dried samples were reconstituted with 650 μ l 0.1M $^2\text{H}_2\text{O}$ based phosphate buffer, pH 7, containing 0.5 mM 3-(trimethylsilyl)-2,2',3,3'-tetrauteropropionic acid (TSP) and 1 g/l NaN_3 . To guarantee solubilisation of the pellet the samples were sonicated for 5 minutes and vortexed. Samples were centrifuged for 5 minutes at 13,000 g to separate insoluble debris from the supernatant, of which 600 μ l were transferred into a 5 mm NMR tube.

B3.2. Acquisition

NMR acquisition was performed as described previously [37].

As most experiments were conducted on different magnets and probes, using different pulse sequences, the exact protocols will be described in the corresponding method section of each chapter.

B3.3. Processing and analysis

Also the NMR processing regime was dependant on the experiment and can be found in each experiment's methods section.

B4. Gas Chromatography - Mass Spectrometry

All GC-MS acquisitions were performed according to a standardised protocol by Kind *et al.* ("The Fiehn Method") [323], in order to use their retention time locked fragmentation library for identification. Their derivatisation procedure consists of a two step protocol with methoximation, followed by silylation.

B4.1. Preparation

10 μ l of 1.5 mg/ml 27-d-myristic acid (in methanol) were added to the sample for retention time locking. Samples must then be dried thoroughly in a vacuum concentrator prior to derivatisation. In the first step, samples were incubated with 20 μ l of 20 mg/ml methoxyamine hydrochloride in anhydrous pyridine for 90 minutes at 30°C. For the second step, samples were silylated with 90 μ l of N-Methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) for 30 minutes at 37°C. During both incubation steps, vials were briefly agitated every 15 minutes to ensure mixing of sample and the derivatisation reagents. Finally, 10 μ l of 1mM 2-fluorobiphenyl in anhydrous pyridine was added as injection standard. Using a glass pipette, the derivatised sample was transferred into silanised high recovery vial inserts (Agilent), which itself was placed back into the original vials.

B4.2. Acquisition

For the acquisition an Agilent 6890 (Wokingham, UK) gas chromatograph with split/splitless injector was used. 1 μ l was injected into a glass-wool split liner at 250°C and a split ratio of 1:5-1:10, with a 3-10 ml/min helium gas flow. A DB5-MS column (30 m length, 0.25 mm inner diameter, 0.25 μ m film, 95% dimethyl 5% diphenyl polysiloxane) was used. The oven was operated at a 1 ml/min constant flow of helium, with a temperature ramp from

60°C (one minute hold) to 325°C at 10°C/min. A 10 minute hold was performed before cool down to 70°C, leading to a 37.5 min total runtime per sample.

As detector, an Agilent 5973 MSD single quadrupole mass spectrometer with electron impact ionisation (EI) at 70 eV was used. The transfer line was kept at 290°C, the filament source at 290°C and the quadrupole at 250°C, whilst scanning over a range of 50-600 u at 2 spectra/s. Fragmentation and acquisition started after a 5.90 min solvent delay.

Samples were handled by an Agilent 7693 autosampler. The whole setup was controlled by Agilent's Chemstation software.

B4.3. Processing

Acquired spectra were converted into the netCDF format before using the Automated Mass Spectral Deconvolution and Identification System (AMDIS) [539], available free of charge from the National Institute of Standards and Technology (NIST) (Gaithersburg, MD, USA) for deconvolution. Settings for peak detection and matching to the Fiehn retention time locked fragmentation database [323] or authentic standards were optimised for each sample set. However, those were kept conservative with high retention time and fragmentation deviation bias in order to minimise false positives [540]. As AMDIS does not integrate the baseline if a fragmentation component was not detected inside a sample, baseline re-integration for each metabolite was performed using the Matlab based software add-on GAVIN [38]. This led to a sample $\times$ metabolite intensity matrix.

B4.4. Analysis

The GC-MS data analysis is described in each experiment's method section.

B5. Liquid Chromatography - Mass Spectrometry

As all LC-MS experiments were performed using different instrumentation, parameter sets and processing workflows, experimental details will be discussed in each chapter's method section.

Supplementary tables and figures

C1. Supplementary tables and figures for chapter II

Name	PubChem Key	GC-MS	NMR	Class
Betaine	247		x	Biogenic amine
Choline	305		x	Biogenic amine
O-phosphocolamine	1015	x		Biogenic amine
Phosphocholine	1014		x	Biogenic amine
putrescine	1045	x		Biogenic amine
3-aminoisobutyric acid	64956	x		Amino acid
Alanine	602		x	Amino acid
Arginine	232		x	Amino acid
Asparagine	236		x	Amino acid
Aspartate	424		x	Amino acid
Glutamate	611		x	Amino acid
Glutamine	738		x	Amino acid
Glycine	750	x	x	Amino acid
Isoleucine	791	x	x	Amino acid
Leucine	857		x	Amino acid
Lysine	866		x	Amino acid
Methionine	876	x		Amino acid
Phenylalanine	994	x	x	Amino acid
Proline	614	x		Amino acid
Pyroglutamic acid	7405	x		Amino acid
Serine	617	x	x	Amino acid
Threonine	205	x		Amino acid
Threonine	205		x	Amino acid
Tyrosine	1153		x	Amino acid
Valine	1182	x	x	Amino acid
Beta- alanine	239	x		Amino acid Derivative/Precursor
Cystathionine	834		x	Amino acid Derivative/Precursor
D-Ala-D-Ala	601	x		Amino acid Derivative/Precursor
Homoserine	779	x		Amino acid Derivative/Precursor
Hydroxy-proline	5810	x		Amino acid Derivative/Precursor
Ornithine	389	x		Amino acid Derivative/Precursor
D-(+)-Trehalose	1143	x	x	Carbohydrate
Glucose	5793		x	Carbohydrate
3-Glycerophosphate	754	x		Carbohydrate Derivative/Precursor
3-Phosphoglycerate	724	x		Carbohydrate Derivative/Precursor
D-glucose-6-phosphate	5958	x		Carbohydrate Derivative/Precursor
Gluconic acid	604	x		Carbohydrate Derivative/Precursor
Glycerol	753	x	x	Carbohydrate Derivative/Precursor
myo-Inositol	892	x		Carbohydrate Derivative/Precursor
sn-Glycero-3-phosphocholine	71920		x	Carbohydrate Derivative/Precursor
Sucrose	5988	x		Carbohydrate Derivative/Precursor
Acetate	175		x	Carboxylic acid
Citramalic acid	1081	x		Carboxylic acid
Citric acid	311	x		Carboxylic acid
Formate	283		x	Carboxylic acid
Fumarate	723	x	x	Carboxylic acid
Lactate	612	x	x	Carboxylic acid
Malate	525	x		Carboxylic acid
Phosphoenolpyruvic acid	1005	x		Carboxylic acid
Phosphoglycolic acid	529	x		Carboxylic acid
Propionate	1032		x	Carboxylic acid
Succinate	1110	x	x	Carboxylic acid
Heptadecanoic acid	10456	x		Fatty acid
Linoleic acid	3931	x		Fatty acid
Palmitic acid	985	x		Fatty acid
Stearic acid	5281	x		Fatty acid
Adenosine	191	x		Nucleoside
Guanosine	765	x		Nucleoside
Xanthosine	1189	x		Nucleoside
Adenine	190	x		Nucleoside base
Thymine	1135	x		Nucleoside base
Uracil	1174	x		Nucleoside base
Xanthine	1188	x		Nucleoside base
AMP	224		x	Nucleotide
ATP	238		x	Nucleotide
GTP	762		x	Nucleotide
Inosine 5-monophosphate	8582	x		Nucleotide
UMP	6030	x		Nucleotide
Pyrophosphate	1023	x		Phosphate
Nicotinic acid	938	x		Vitamin or derivative
Pantothenic acid	988	x		Vitamin or derivative
Cholesterol	304	x		Steroid
Methoxytryptamine	1833	x		Biogenic amine
Phenylethanolamine	1000	x		Biogenic amine

Table C1.1.: *C. elegans* compounds detected via GC-MS and NMR:
Pubchem identifier and compound class is also given.

C2. Supplementary tables and figures for chapter III

Table C2.1.: **HMDB metabolites matched with ^{12}C and ^{13}C HSQC cross peaks:**
Metabolites returned by searching the ^{12}C and ^{13}C HSQC cross peaks against the HMDB HSQC reference spectra. Table contains name, HMDB[630] and PubChem[56] identifier, the number of cross peaks for each metabolite and the % fraction which of those were detected in the ^{12}C and ^{13}C HSQC spectrum. If no number is given, no peak was detected.

No	Metabolite Name	HMDB id	PubChem id	Peaks	% ^{13}C	% ^{12}C
1	Citicoline	HMDB01413	13804	11	100	55
2	D-Glucose	HMDB00122	5793	10	100	30
3	Glycogen	HMDB00757	439177	10	100	25
5	Trehalose	HMDB00975	7427	8	100	100
4	Glycerophosphocholine	HMDB00086	71920	8	100	
6	Cytidine triphosphate	HMDB00082	6176	7	100	43
7	Spermidine	HMDB01257	1102	7	100	29
8	Guanosine diphosphate	HMDB01201	8977	6	100	67
10	Glycerol 3-phosphate	HMDB00126	439162	5	100	100
16	D-Arginine	HMDB03416	71070	5	100	60
12	Ornithine	HMDB00214	6262	5	100	60
9	L-Cystathionine	HMDB00099	439258	5	100	40
15	Spermine	HMDB01256	1103	5	100	40
11	L-Lysine	HMDB00182	5962	5	100	20
13	Glycyl-L-leucine	HMDB00759	92843	5	100	
14	N-Acetylglutamic acid	HMDB01138	185	5	100	
22	D-Xylitol	HMDB02917	6912	4	100	100
18	L-Arginine	HMDB00517	6322	4	100	75
21	Threonic acid	HMDB00943	151152	4	100	75
17	Ribitol	HMDB00508	6912	4	100	50
19	L-2-Hydroxyglutaric acid	HMDB00694	439939	4	100	25
20	L-Valine	HMDB00883	6287	4	100	25
24	L-Glutamic acid	HMDB00148	33032	3	100	100
29	3-Phosphoglyceric acid	HMDB00807	724	3	100	67
23	Glycerol	HMDB00131	753	3	100	67
25	L-Alanine	HMDB00161	5950	3	100	67
27	L-Glutamine	HMDB00641	5961	3	100	67
31	D-Glutamine	HMDB03423	145815	3	100	33
32	D-Leucine	HMDB13773	6106	3	100	
26	L-Asparagine	HMDB00168	6267	3	100	
28	L-Leucine	HMDB00687	6106	3	100	

Continued on next page

Table C2.1 – Continued from previous page

No	Metabolite Name	HMDB id	PubChem id	Peaks	% 13C	% 12C
30	N-Acetyl-L-aspartic acid	HMDB00812	65065	3	100	
40	Acetylglycine	HMDB00532	10972	2	100	50
44	D-Lactic acid	HMDB01311	61503	2	100	50
37	L-Lactic acid	HMDB00190	107689	2	100	50
36	L-Serine	HMDB00187	5951	2	100	50
45	Putrescine	HMDB01414	1045	2	100	50
46	Acetylphosphate	HMDB01494	186	2	100	
33	Beta-Alanine	HMDB00056	239	2	100	
34	Citric acid	HMDB00094	311	2	100	
43	D-Alanine	HMDB01310	71080	2	100	
35	Ethanolamine	HMDB00149	700	2	100	
41	Glutaric acid	HMDB00661	743	2	100	
42	Hydroxypropionic acid	HMDB00700	68152	2	100	
38	Oxoglutaric acid	HMDB00208	51	2	100	
47	Pectin	HMDB03402	441476	2	100	
39	Phosphoenolpyruvic acid	HMDB00263	1005	2	100	
50	Glycine	HMDB00123	750	1	100	100
52	Succinic acid	HMDB00254	1110	1	100	100
54	Tartaric acid	HMDB00956	444305	1	100	100
48	Acetic acid	HMDB00042	176	1	100	
55	Dihydroxyacetone	HMDB01882	670	1	100	
57	Dimethyl sulfone	HMDB04983	6213	1	100	
49	Dimethylamine	HMDB00087	674	1	100	
56	Dimethylmalonic acid	HMDB02001	11686	1	100	
53	Malonic acid	HMDB00691	867	1	100	
51	Orotic acid	HMDB00226	967	1	100	
58	Uridine diphosphate-N-acetylglucosamine	HMDB00290	445675	15	93	73
59	Maltotetraose	HMDB01296	439639	18	89	39
60	D-Glucuronic acid	HMDB00127	444791	9	89	33
61	D-Maltose	HMDB00163	10991489	17	88	35
63	Adenosine triphosphate	HMDB00538	5957	8	88	75
62	Allocystathionine	HMDB00455	10104953	8	88	25
65	Adenosine 3',5'-diphosphate	HMDB00061	159296	7	86	71
70	Adenosine phosphosulfate	HMDB01003	10238	7	86	71
71	Guanosine triphosphate	HMDB01273	6830	7	86	71
64	Adenosine monophosphate	HMDB00045	6083	7	86	57
68	Uridine 5'-diphosphate	HMDB00295	6031	7	86	57
66	L-Aspartic acid	HMDB00191	5960	7	86	
69	N-Alpha-acetyllysine	HMDB00446	192590	7	86	

Continued on next page

Table C2.1 – Continued from previous page

No	Metabolite Name	HMDB id	PubChem id	Peaks	% 13C	% 12C
67	Pantothenic acid	HMDB00210	988	7	86	
72	Alpha-D-Glucose	HMDB03345	79025	13	85	46
75	Guanosine monophosphate	HMDB01397	6804	6	83	50
74	L-Isoleucine	HMDB00172	6306	6	83	
73	L-Phenylalanine	HMDB00159	6140	6	83	
76	Uridine diphosphate glucose	HMDB00286	53477679	17	82	65
81	D-Lysine	HMDB03405	57449	5	80	40
78	Gamma-Glutamylcysteine	HMDB01049	123938	5	80	40
77	Homo-L-arginine	HMDB00670	9085	5	80	20
79	Aminocaproic acid	HMDB01901	564	5	80	
80	Erythrose	HMDB02649	439574	5	80	
82	Maltitol	HMDB02928	493591	14	79	57
83	N1-Acetylspermine	HMDB01186	916	9	78	22
84	Uridine diphosphategalactose	HMDB00302	18068	13	77	54
85	Inosinic acid	HMDB00175	8582	8	75	63
86	Sorbitol	HMDB00247	5780	8	75	63
87	1-Methylguanosine	HMDB01563	96373	8	75	
88	Alpha-Aspartyl-lysine	HMDB04987	6427003	8	75	
92	L-Iditol	HMDB11632	5460044	4	75	75
89	(R)-3-Hydroxyisobutyric acid	HMDB00336	11217234	4	75	
91	3-Aminoisobutanoic acid	HMDB03911	64956	4	75	
90	Citrulline	HMDB00904	9750	4	75	
94	Beta-N-Acetylglucosamine	HMDB00803	24139	15	73	60
93	N-Acetyl-D-glucosamine	HMDB00215	11861101	15	73	60
95	Maltotriose	HMDB01262	439586	18	72	39
97	Glucose 1-phosphate	HMDB01586	439165	7	71	43
96	Isovalerylglutamic acid	HMDB00726	133383	7	71	
98	NAD	HMDB00902	5893	19	68	32
99	D-Xylose	HMDB00098	135191	18	67	22
100	7-Methylguanosine	HMDB01107	445404	9	67	22
102	Glutathione	HMDB00125	124886	6	67	50
106	Gluconic acid	HMDB00625	10690	6	67	17
104	Gluconolactone	HMDB00150	7027	6	67	17
105	N6-Acetyl-L-lysine	HMDB00206	92832	6	67	17
103	Guanosine	HMDB00133	6802	6	67	
101	Iodotyrosine	HMDB00021	439744	6	67	
107	N-Acetyl-L-tyrosine	HMDB00866	68310	6	67	
121	Phosphorylcholine	HMDB01565	1014	3	67	100
128	D-Threitol	HMDB04136	169019	3	67	67

Continued on next page

Table C2.1 – Continued from previous page

No	Metabolite Name	HMDB id	PubChem id	Peaks	% 13C	% 12C
108	L-Carnitine	HMDB00062	2724480	3	67	67
110	L-Threonine	HMDB00167	6288	3	67	67
125	Canavanine	HMDB02706	439202	3	67	33
119	Diaminopimelic acid	HMDB01370	439283	3	67	33
117	Homocysteine	HMDB00742	778	3	67	33
118	Propionylglycine	HMDB00783	98681	3	67	33
131	3-Dechloroethylfosfamide	HMDB13858	1983	3	67	
120	3-Hydroxyanthranilic acid	HMDB01476	86	3	67	
127	4-Guanidinobutanoic acid	HMDB03464	500	3	67	
129	4-Heptanone	HMDB04814	31246	3	67	
115	4-Hydroxyphenylpyruvic acid	HMDB00707	979	3	67	
123	Acetaminophen	HMDB01859	1983	3	67	
124	Acetylcysteine	HMDB01890	12035	3	67	
113	D-Alpha-aminobutyric acid	HMDB00650	439691	3	67	
122	DL-O-Phosphoserine	HMDB01721	106	3	67	
109	Gamma-Aminobutyric acid	HMDB00112	119	3	67	
114	Ketoleucine	HMDB00695	70	3	67	
112	L-Alpha-aminobutyric acid	HMDB00452	80283	3	67	
126	L-Dihydroorotic acid	HMDB03349	439216	3	67	
116	Levulinic acid	HMDB00720	11579	3	67	
111	Phenol	HMDB00228	996	3	67	
130	Sulfamethoxazole N1-glucuronide	HMDB13855	996	3	67	
134	Glucose 6-phosphate	HMDB01401	5958	11	64	36
133	N-Acetylmannosamine	HMDB01129	11096158	11	64	36
132	NADP	HMDB00217	5886	11	64	27
139	Uridine 5'-monophosphate	HMDB00288	6030	8	63	63
137	D-Mannose	HMDB00169	18950	8	63	50
135	Argininosuccinic acid	HMDB00052	16950	8	63	38
138	Saccharopine	HMDB00279	160556	8	63	38
136	Cytidine monophosphate	HMDB00095	6131	8	63	25
140	Cysteineglutathione disulfide	HMDB00656	53477713	10	60	30
146	Hydroxyphenyllactic acid	HMDB00755	9378	5	60	40
149	L-Arabitol	HMDB01851	439255	5	60	40
143	2-Hydroxyadipic acid	HMDB00321	193530	5	60	20
151	3-Phenoxypropionic acid	HMDB02229	81596	5	60	20
148	6-Phosphogluconic acid	HMDB01316	91493	5	60	20
145	Homocitrulline	HMDB00679	65072	5	60	20
142	L-Tyrosine	HMDB00158	6057	5	60	20
150	N-Acetylputrescine	HMDB02064	122356	5	60	20

Continued on next page

Table C2.1 – Continued from previous page

No	Metabolite Name	HMDB id	PubChem id	Peaks	% 13C	% 12C
141	(S)-3-Hydroxyisobutyric acid	HMDB00023	440873	5	60	
144	Leucinic acid	HMDB00665	92779	5	60	
147	Ureidosuccinic acid	HMDB00828	93072	5	60	
152	N-Acetylgalactosamine 4-sulphate	HMDB00781	446101	12	58	33
153	Allose	HMDB01151	12285879	7	57	29
154	5-Phenylvaleric acid	HMDB02043	16757	7	57	
155	Coenzyme A	HMDB01423	6816	16	56	25
156	Inosine triphosphate	HMDB00189	8583	9	56	56
159	Mannose 6-phosphate	HMDB01078	439198	9	56	33
157	Aminoadipic acid	HMDB00510	469	9	56	11
160	D-Pantethine	HMDB03828	452306	9	56	
158	L-Aspartyl-L-phenylalanine	HMDB00706	93078	9	56	
161	Cellobiose	HMDB00055	10712	20	55	45
162	Alpha-Lactose	HMDB00186	84571	20	55	30
164	Furan	HMDB13785	5988	11	55	27
163	Sucrose	HMDB00258	5988	11	55	27
165	L-Arabinose	HMDB00646	439195	12	50	25
166	Homocarnosine	HMDB00745	10243361	8	50	13
167	5-Methoxydimethyltryptamine	HMDB02004	1832	8	50	
177	D-Xylulose	HMDB01644	5289590	6	50	50
170	L-Acetylcarnitine	HMDB00201	1	6	50	17
171	2-Methyl-3-ketovaleric acid	HMDB00408	189028	6	50	
168	3-Methoxytyramine	HMDB00022	1669	6	50	
173	3-Methyl-2-oxovaleric acid	HMDB00491	47	6	50	
172	3,4-Dihydroxyhydrocinnamic acid	HMDB00423	348154	6	50	
174	DL-Dopa	HMDB00609	836	6	50	
169	L-Dopa	HMDB00181	6047	6	50	
176	N-Formyl-L-methionine	HMDB01015	6995182	6	50	
175	Valerylglycine	HMDB00927	4737557	6	50	
195	2,3-Butanediol	HMDB03156	262	4	50	25
181	Glutarylglycine	HMDB00590	23592950	4	50	25
183	Isovalerylglycine	HMDB00678	546304	4	50	25
185	L-Homoserine	HMDB00719	12647	4	50	25
184	L-Methionine	HMDB00696	6137	4	50	25
192	Methionine sulfoxide	HMDB02005	847	4	50	25
193	N-Methyl-D-aspartic acid	HMDB02393	22880	4	50	25
187	Neopterin	HMDB00845	4455	4	50	25
186	Pyrrolidonecarboxylic acid	HMDB00805	499	4	50	25
179	Tyramine	HMDB00306	5610	4	50	25

Continued on next page

Table C2.1 – Continued from previous page

No	Metabolite Name	HMDB id	PubChem id	Peaks	% 13C	% 12C
190	2-Ketohexanoic acid	HMDB01864	159664	4	50	
180	4,5-Dihydroorotic acid	HMDB00528	648	4	50	
196	5-Aminopentanoic acid	HMDB03355	138	4	50	
191	D-threo-Isocitric acid	HMDB01874	5318532	4	50	
188	Histamine	HMDB00870	774	4	50	
194	Methylimidazoleacetic acid	HMDB02820	75810	4	50	
178	Phenylacetic acid	HMDB00209	999	4	50	
182	Succinylacetone	HMDB00635	5312	4	50	
189	trans-Cinnamic acid	HMDB00930	444539	4	50	
200	Betaine	HMDB00043	247	2	50	100
202	Creatine	HMDB00064	586	2	50	50
219	N-Methyl- α -aminoisobutyric acid	HMDB02141	6951124	2	50	50
214	Phosphocreatine	HMDB01511	587	2	50	50
197	1,3-Diaminopropane	HMDB00002	428	2	50	
198	2-Ketobutyric acid	HMDB00005	58	2	50	
217	2,2-Dimethylsuccinic acid	HMDB02074	11701	2	50	
209	5-Hydroxymethyluracil	HMDB00469	78168	2	50	
220	A-Ketoglutaric acid oxime	HMDB02467	13012726	2	50	
201	Acetoacetic acid	HMDB00060	96	2	50	
215	Acetone	HMDB01659	180	2	50	
208	Adipic acid	HMDB00448	196	2	50	
222	Diethanolamine	HMDB04437	8113	2	50	
203	Dihydrouracil	HMDB00076	649	2	50	
204	Dimethylglycine	HMDB00092	673	2	50	
211	Galactaric acid	HMDB00639	3037582	2	50	
216	Isobutyric acid	HMDB01873	6590	2	50	
212	Isopropyl alcohol	HMDB00863	3776	2	50	
210	L-Cysteine	HMDB00574	5862	2	50	
221	Phosphonoacetate	HMDB04110	546	2	50	
205	Propionic acid	HMDB00237	1032	2	50	
213	Pyrocatechol	HMDB00957	289	2	50	
207	Sarcosine	HMDB00271	1088	2	50	
218	Syringic acid	HMDB02085	10742	2	50	
206	Taurine	HMDB00251	1123	2	50	
199	Ureidopropionic acid	HMDB00026	111	2	50	
223	5-Hydroxylysine	HMDB00450	3032849	9	44	33
224	dATP	HMDB01532	15993	9	44	22
232	D-Tagatose	HMDB03418	92092	7	43	43
233	5-Methoxytryptamine	HMDB04095	1833	7	43	14

Continued on next page

Table C2.1 – Continued from previous page

No	Metabolite Name	HMDB id	PubChem id	Peaks	% 13C	% 12C
225	Carnosine	HMDB00033	439224	7	43	14
227	Glucosan	HMDB00640	2724705	7	43	14
228	Suberylglycine	HMDB00953	6453952	7	43	14
229	2'-Deoxyguanosine 5'-monophosphate	HMDB01044	65059	7	43	
230	dGTP	HMDB01440	65103	7	43	
231	Diphenhydramine	HMDB01927	3100	7	43	
234	o-Tyrosine	HMDB06050	91482	7	43	
226	Xanthosine	HMDB00299	64959	7	43	
236	Glucosamine	HMDB01514	45933887	12	42	25
235	2-Methylbutyrylcarnitine	HMDB00378	6426901	12	42	17
237	Lactulose	HMDB00740	11024312	22	41	14
239	3,4-Dihydroxyphenylglycol	HMDB00318	91528	5	40	20
241	D-Arabitol	HMDB00568	827	5	40	20
247	L-Norleucine	HMDB01645	21236	5	40	20
242	Ortho-Hydroxyphenylacetic acid	HMDB00669	11970	5	40	20
245	Phenylacetyl glycine	HMDB00821	68144	5	40	20
240	2-Hydroxy-3-methylbutyric acid	HMDB00407	99823	5	40	
238	Dopamine	HMDB00073	681	5	40	
244	Hydrocinnamic acid	HMDB00764	107	5	40	
243	Hydroxyisocaproic acid	HMDB00746	83697	5	40	
246	trans-Ferulic acid	HMDB00954	445858	5	40	
248	Stachyose	HMDB03553	439531	23	39	22
249	NADH	HMDB01487	928	18	39	28
253	Ribothymidine	HMDB00884	445408	8	38	38
251	Uridine	HMDB00296	6029	8	38	38
255	5-Methylcytidine	HMDB00982	92918	8	38	25
250	Cytidine	HMDB00089	6175	8	38	25
256	dCMP	HMDB01202	13945	8	38	13
252	Isobutyryl-L-carnitine	HMDB00736	10177002	8	38	
254	L-Tryptophan	HMDB00929	6305	8	38	
257	D-Galactose	HMDB00143	439357	11	36	9
258	L(-)-Nicotine pestanal	HMDB01934	89594	11	36	
259	1-Methyladenosine	HMDB03331	27476	17	35	12
260	Pteroyltriglutamic acid	HMDB01902	6968	15	33	
261	Metoprolol	HMDB01932	4171	12	33	8
262	Fructose 6-phosphate	HMDB00124	69507	9	33	22
263	Inosine	HMDB00195	6021	9	33	11
267	3-Methylhistidine	HMDB00479	64969	6	33	17
265	L-Histidine	HMDB00177	6274	6	33	17

Continued on next page

Table C2.1 – *Continued from previous page*

No	Metabolite Name	HMDB id	PubChem id	Peaks	% 13C	% 12C
266	Serotonin	HMDB00259	5202	6	33	17
271	3-Chlorotyrosine	HMDB01885	119226	6	33	
272	3-Phenylbutyric acid	HMDB01955	20724	6	33	
264	Epinephrine	HMDB00068	5816	6	33	
269	L-3-Phenyllactic acid	HMDB00748	643327	6	33	
268	L-Kynurenine	HMDB00684	161166	6	33	
270	Phenyllactic acid	HMDB00779	3848	6	33	
288	2,3-Diaminopropionic acid	HMDB02006	97328	3	33	67
278	6-Dimethylaminopurine	HMDB00473	3134	3	33	33
274	Choline	HMDB00097	305	3	33	33
294	Erythritol	HMDB02994	222285	3	33	33
275	Galactitol	HMDB00107	11850	3	33	33
273	p-Hydroxyphenylacetic acid	HMDB00020	127	3	33	33
292	1,11-Undecanedicarboxylic acid	HMDB02327	10458	3	33	
289	2-Pyrrolidinone	HMDB02039	12025	3	33	
293	2,4-Diaminobutyric acid	HMDB02362	470	3	33	
301	2,4-Dichlorophenol	HMDB04811	8449	3	33	
286	5-Aminolevulinic acid	HMDB01149	137	3	33	
296	Acetoin	HMDB03243	179	3	33	
291	Cadaverine	HMDB02322	273	3	33	
302	cis-4-Octenedioic acid	HMDB04982	11805205	3	33	
297	Cyclohexanone	HMDB03315	7967	3	33	
281	Cysteine-S-sulfate	HMDB00731	115015	3	33	
298	D-Cysteine	HMDB03417	92851	3	33	
279	Gamma-Butyrolactone	HMDB00549	7302	3	33	
295	Guanidinosuccinic acid	HMDB03157	439918	3	33	
280	Isovaleric acid	HMDB00718	10430	3	33	
290	Itaconic acid	HMDB02092	811	3	33	
299	L-Allothreonine	HMDB04041	99289	3	33	
277	L-Cystine	HMDB00192	67678	3	33	
276	L-Malic acid	HMDB00156	222656	3	33	
282	Malic acid	HMDB00744	525	3	33	
283	Pimelic acid	HMDB00857	385	3	33	
300	Selenocystine	HMDB04122	15104	3	33	
285	Suberic acid	HMDB00893	10457	3	33	
287	Theophylline	HMDB01889	2153	3	33	
284	Undecanedioic acid	HMDB00888	15816	3	33	

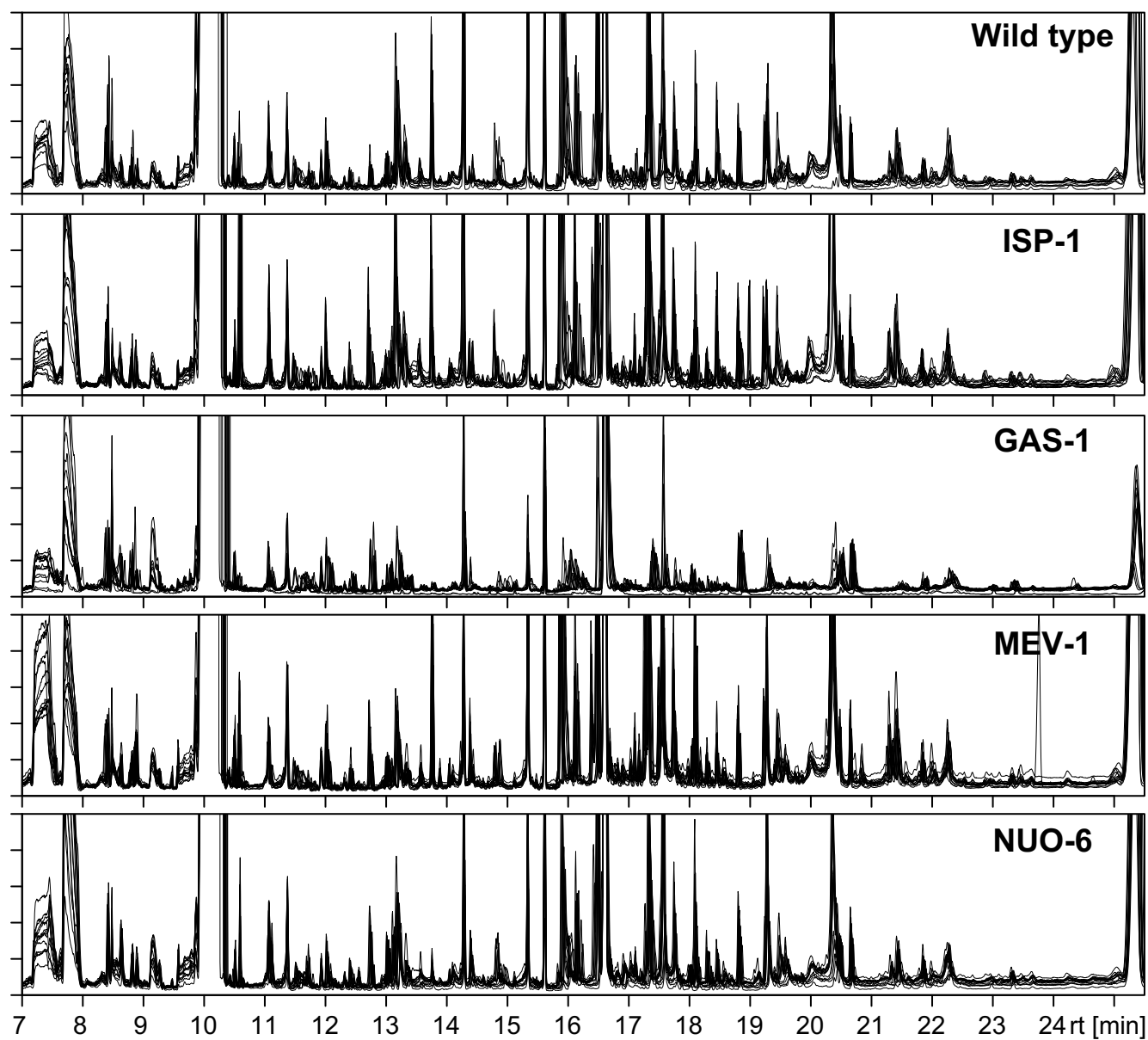


Figure C2.1.: Aligned spectra of the mitochondrial mutant dataset:
The normalised total ion chromatograms show clear differences in

C3. Supplementary tables and figures for chapter IV

Element	minimum [# Atoms/Da]	maximum[# $\hat{E}$ Atoms/Da]
Carbon	0.011	0.080
Nitrogen	0	0.040
Sulphur	0	0.018

Table C3.1.: **Minimum and maximum atom-count to m/z ratios:**

Determined as 0.01% and 99.99% PubChem confidence intervals

Element	minimum [# $\hat{E}$ Carbon/ # Element]	maximum[# Carbon/ # Element]
Hydrogen	0.08	7.00
Nitrogen	0	5.00
Oxygen	0	6.00
Phosphorous	0	3.00
Sulphur	0	2.5

Table C3.2.: **Minimum and maximum elemental ratios:**

Determined as 0.01% and 99.99% PubChem confidence intervals. Carbon is used as the reference, as any molecule is required to have at least one carbon atom.

Adduct	Charge/Mode
M-H	1-
M-H ₂ O-H	1-
M+Na-2H	1-
M+Cl	1-
M+K-2H	1-
M+FA-H	1-
M+Hac-H	1-
M+Br	1-
2M-H	1-
2M+FA-H	1-
2M+Hac-H	1-
3M-H	1-
M-2H	2-
M-3H	3-
M+H	1+
M+NH ₄	1+
M+Na	1+
M+CH ₃ OH+H	1+
M+K	1+
M+ACN+H	1+
M+2Na-H	1+
M+ACN+Na	1+
M+2K+H	1+
M+2ACN+H	1+
2M+H	1+
2M+NH ₄	1+
2M+Na	1+
2M+K	1+
2M+ACN+H	1+
2M+ACN+Na	1+
M+	1+
M+2H	2+
M+H+NH ₄	2+
M+H+Na	2+
M+H+K	2+
M+ACN+2H	2+
M+2Na	2+
M+2ACN+2H	2+
M+3ACN+2H	2+
2M+3H ₂ O+2H	2+
M+3H	3+
M+2H+Na	3+
M+H+2Na	3+
M+3Na	3+

Table C3.3.: Adducts used during molecular formula calculation.

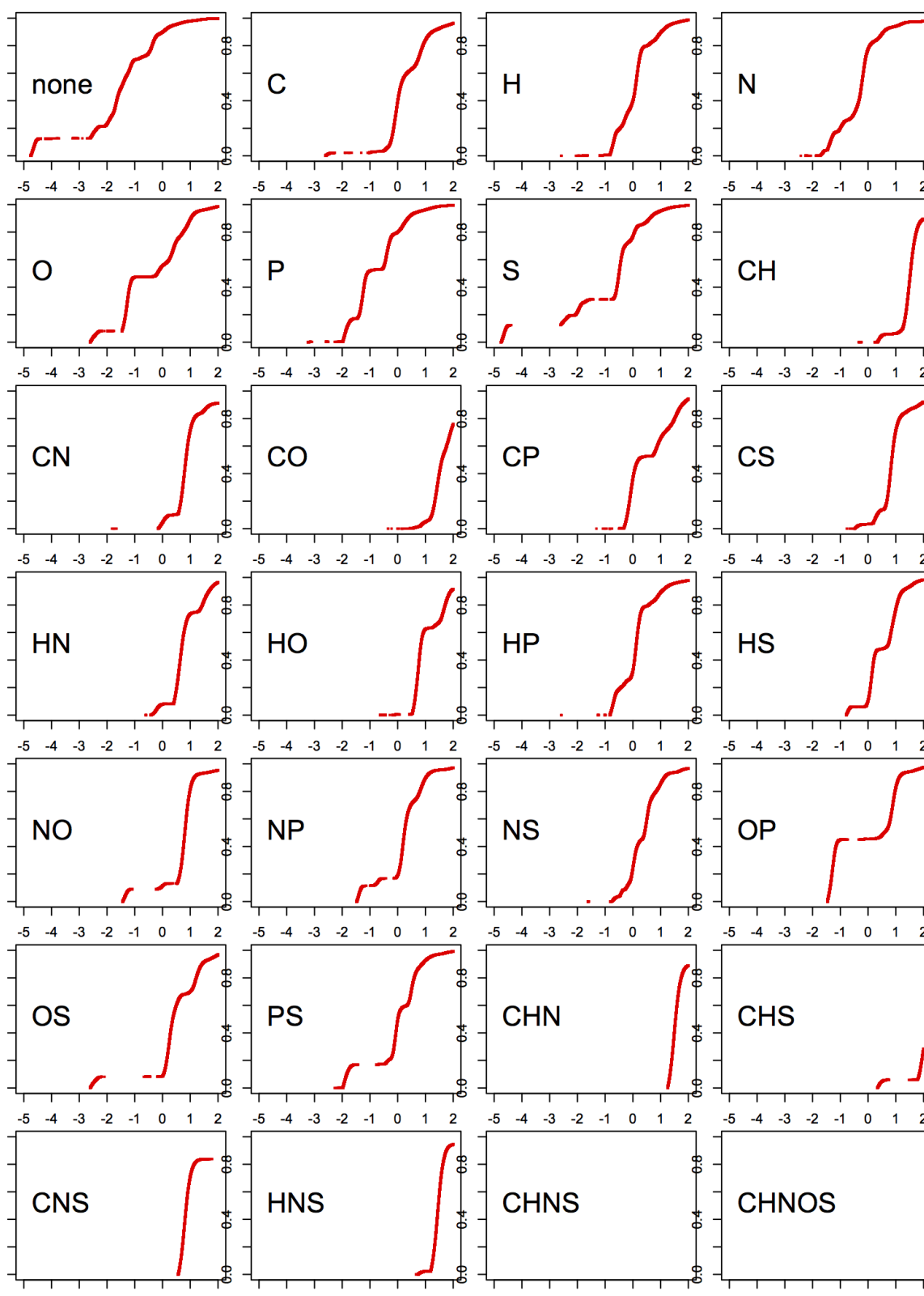


Figure C3.1.: Accuracy requirement for *ab initio* formula calculation, impacted by various atom-count constraints:
x-axis is $\log_{10}(\text{ppm})$, y-axis is yet unresolved fraction of metabolites. Refer to chapter IV.3.1 for more information.

C4. Supplementary tables and figures for chapter V

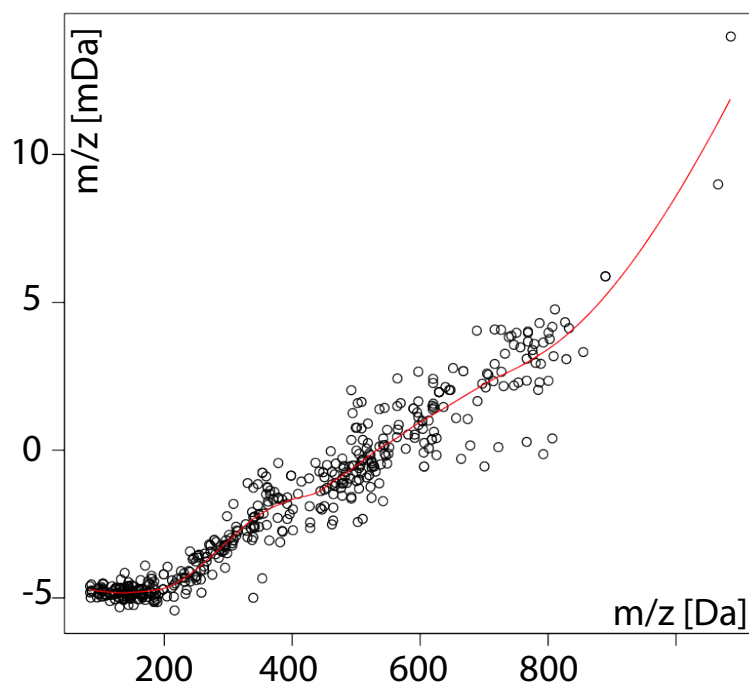


Figure C4.1.: **LOESS function to correct post-acquisition calibrate m/z domain:**

The individual data points are the difference (in measured mass) of the same compound before and after the power outage. All measurements across the whole mass range show a clear trend and therefore allowing post-acquisition correction of the m/z domain (for the post power outage samples). In red, a LOESS (locally weighted scatterplot smoothing) regression function is fitted through the data points. This function can then be used as a calibration curve, allowing to correct the masses of the affected samples.

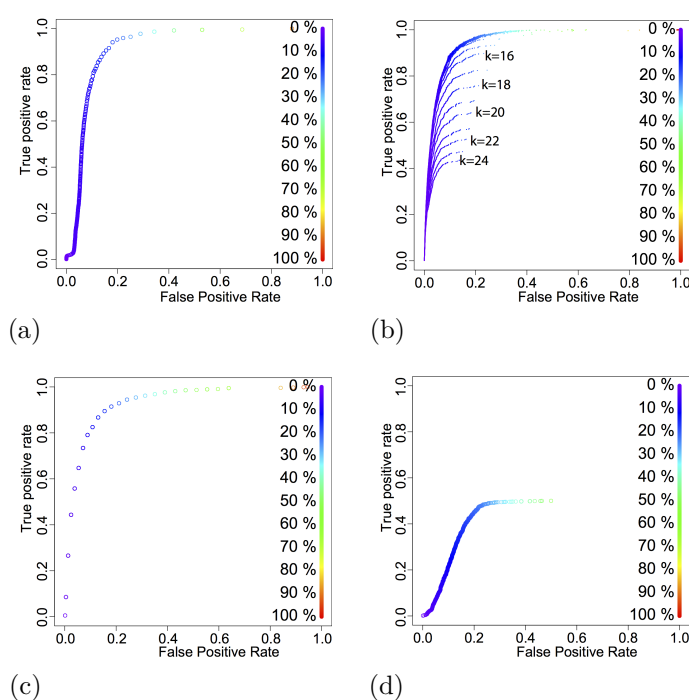


Figure C4.2.: **ROC curves for optimising xcms peak detection parameters:**

Receiver operating characteristics (ROC) analysis was used to try optimise xcms peak detection parameters, to potentially cut the numbers of detected features. For more details, please refer to the appropriate methods or results section V.2.3.3 and V.3.3, respectively. Figure C4.2a shows the ROC curve for the *SN*, figure C4.2b for the *prefilter*, figure C4.2c for the *PPM* and figure C4.2d the *peakwidth* parameter. The colour indicates by how much a certain parameter threshold would reduce the number of detected data points.

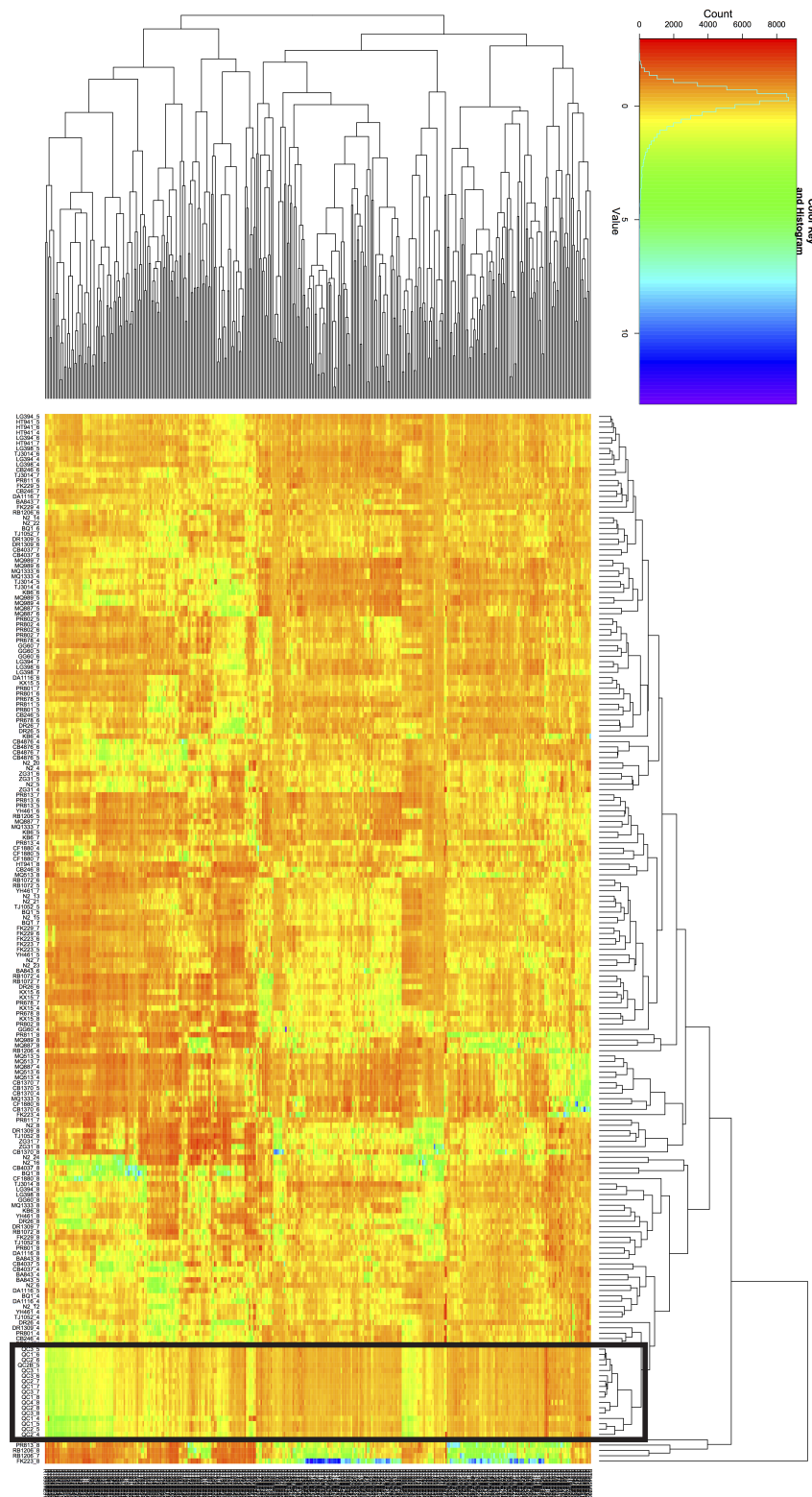


Figure C4.3.: **Hierarchical clustered heat map of all samples:**

All samples are unit variance scaled. The cluster of QC samples is highlighted by a black square. Within all samples, patterns are visible. However, the data structure appears to be highly complex.

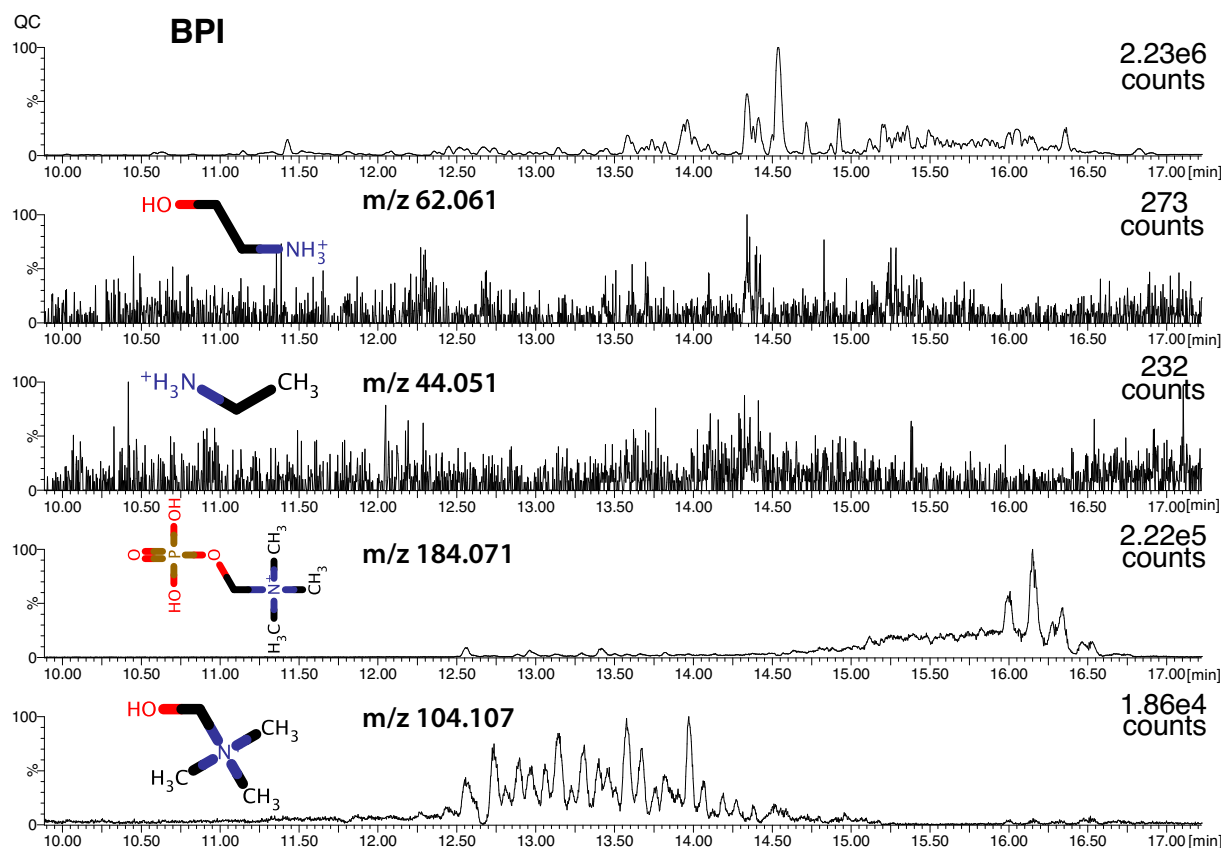


Figure C4.4.: **Glycerophosphocholines versus glycerophosphoethanolamines:**

This figure shows the base ion peak (BPI) chromatogram and several extracted ion chromatograms (EIC) of the lipid region of a QC sample. It can be seen, that there is little evidence for ethanolamine or aminoethane, which would indicate the presence of glycerophosphoethanolamines. Instead, choline and glycerophosphocholine fragments, hinting at glycerophosphocholines can be seen. The earlier eluting monoacylglycerophosphocholines seem to preferentially lose phosphocholine, whereas the later eluting diacylglycerophosphocholines are more likely to only split of choline.