Design and Implementation of Efficient Workflows for Computational Metabolomics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Inform. (Bioinformatik) Erhan Kenar
aus Tuttlingen

Tübingen
2015

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	15.06.15
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Rainer Lehmann

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisors Oliver Kohlbacher and Rainer Lehmann for giving me the opportunity to undertake this PhD project. Throughout the course of my graduate studies, their guidance and support gave my the confidence to realize this work. I am especially thankful for Oliver always having an open ear for my concerns.

Furthermore, I would like to thank my fellow colleagues from the *Applied Bioinformatics* group for fruitful discussions in an enjoyable atmosphere and relaxing lunch breaks in the Mensa: Fabian Aicheler, Linus Backert, Thorsten Blum, Sebastian Briesemeister, Magdalena Feldhahn, Nina Fischer, Andreas Friedrich, Luis de la Garza, Sandra Gesing, Andreas Kämper, Jens Krüger, Christopher Mohr, Sven Nahnsen, Peter Niermann, Lars Nilse, Nico Pfeifer, Marc Röttig, Marc Rurik, Timo Sachsenberg, Charlotta Schärfe, Benjamin Schubert, Marcel Schumann, Andras Szolek, Philipp Thiel, Nora Toussaint, Johannes Veit, and Mathias Walzer. Special thanks and warm regards go to Marc Röttig who over the years became a dear friend of mine.

Many thanks go to my colleagues at the Helmholtz-Zentrum München for inspiring discussions and a great collaboration: Agnes Fekete, Sara Forcisi, Kilian Wörmann, and Marianna Lucio. In particular, I am grateful to Philippe Schmitt-Kopplin who invited me on a regular basis to his lab for gaining a deeper understanding of mass spectrometry.

I would like to sincerely thank Rainer Lehmann and Prof. Guowang Xu who enabled my research visit at the CAS Key Laboratory of Separation Science for Analytical Chemistry (Dalian, China) in the course of the Sino German cooperation. Many thanks go to the friendly people from Prof. Xu's lab for the interesting discussions and impulses, in particular, to Xinjie Zhao and Jia Li.

Likewise, I would like to thank Prof. Christian Huber for his kind invitation to his lab at the University of Salzburg. There, I had insightful debates and a fruitful cooperation with Christina

Ranninger and Roland Reischl for which I am grateful. Furthermore, I would like to express my appreciation for Christina's work that provided me with valuable metabolomics data.

I would like to thank Holger Franken for the successful cooperation and amicable conversations.

On a final note, I want to express my deepest thanks to my parents and my whole family who supported me in all my endeavors. My warmest thanks go to Kathrin, for all her love, support, and patience.

Abstract

In recent years, metabolomics has become a powerful approach to systematically study alterations in metabolism induced by disease, nutrition, and environmental changes. Liquid chromatography-mass spectrometry (LC-MS) has been established as the main analytical platform since it is sensitive enough to capture metabolomes' high complexity and chemical diversity and allows for processing biological samples in a high-throughput manner. However, rapid evolvement of mass spectrometry (MS) technology resulted in enormous data volumes and complexity and with that gave rise to a computational bottleneck. Computational metabolomics strives to develop efficient algorithms and flexible workflows to meet the data analysis needs of high-throughput metabolomics experiments. In this work, we addressed the three key problems of computational metabolomics: quantification of metabolites, their identification, and statistical methods to reveal discriminatory metabolic patterns and novel biomarkers. Our main design goal was to develop robust and comprehensive computational workflows to integrate solutions to these key problems consistently instead of addressing them separately.

We developed a novel algorithm for the robust detection and quantification of metabolite features in LC-MS data. It extracts chromatographic profiles with high sensitivity and provides deisotoping by a novel support vector machine (SVM)-based classifier. Our algorithm was validated both on real-world and simulated LC-MS benchmark datasets and showed an excellent performance when compared to existing solutions. In order to identify unknown features, we devised a comprehensive and integrative strategy that exploits as many complementary feature characteristics as possible (e.g., relative isotopic abundance (RIA) and retention time (rt)). These were then combined as filter criteria to yield more reliable metabolite identifications (IDs). To this end, we implemented an accurate mass search which efficiently facilitates any number of queries against a metabolite database. It covers a wide range of potential adducts by default but allows for customization. We augmented it by orthogonal RIA and rt filters to considerably reduce the number of false positive IDs. The utility of rt prediction models has been shown

in a few stray instances, however, they were usually not built and integrated routinely in ID pipelines as in our case. Furthermore, we provided an efficient means of matching tandem mass spectrometry (MS/MS) spectra against a precompiled fragment database. Based on an LC-MS/MS validation dataset, we achieved excellent ID accuracy when our accurate mass search with orthogonal filters and spectral matching algorithms were combined.

The ultimate goal of computational metabolomics is to extract novel biological knowledge from complex LC-MS-based metabolomics data. To this end, we designed comprehensive analysis workflows for untargeted metabolomics data that build on robust statistical methods. We showed their utility by answering two biologically relevant questions. The first question involved the role of the rs7903146 polymorphism located in the transcription factor 7-like 2 (TCF7L2) gene and its impact on type 2 diabetes mellitus (T2DM). We found more than 100 potential biomarkers with some pointing out perturbations in the bile acid and androgenic steroid biosyntheses, two well-documented complications of T2DM. Our approach outperformed classical statistical methods such as the partial least squares discriminant analysis (PLS-DA) which could not detect any significant differences between control and risk allele groups. The second question revolved around finding characteristic kinetic patterns and corresponding metabolic pathways that were perturbed during a two-hour single bout of exercise and a follow-up three-hour recovery stage. We adapted a known clustering algorithm to condense thousands of kinetic profiles that comprise only few and non-equidistant time points, a common scenario in metabolomics time-course experiments. Our novel clustering approach yielded 25 distinct clusters that were characteristic for either exercise or recovery stage. In a pathway enrichment analysis, the two most prominent clusters suggested the involvement of amino acid, free fatty acid (FFA), and catecholamine metabolism. Perturbations of these metabolic pathways were reported before in the context of physical exercise and recovery. Both discovery workflows produced biologically sound results which enabled us to construct more specific hypotheses. Driven by these, we now can validate our results with targeted follow-up experiments while avoiding redundant analyses, reducing time and costs.

Zusammenfassung

In jüngster Vergangenheit hat sich die Metabolomik zu einer leistungsfähigen Methode entwickelt, mittels derer sich systematisch Veränderungen des Stoffwechsels erfassen lassen, welche durch Krankheit, Ernährung oder Umwelteinflüsse herbeigeführt werden. Die Massenspektrometrie mit Flüssigkeitschromatographie-Kopplung hat sich dabei als wichtigstes analytisches Instrument etabliert, da sie sowohl die hohe Komplexität und chemische Vielfalt eines Metaboloms erfassen kann, als auch die Messung biologischer Proben im Hochdurchsatz erlaubt. Die rasante Entwicklung in der Massenspektrometrie führte jedoch zu enormen Datenmengen und erhöhter Datenkomplexität, sodass die rechnerische Auswertung schnell zu einem Engpass führte. Das Gebiet der "Computational Metabolomics" zielt auf die Entwicklung effizienter Algorithmen und flexibler Workflows ab, um den Anforderungen an die Analyse von im Hochdurchsatz gemessenen Metabolomikdaten gerecht zu werden. Im Rahmen dieser Arbeit haben wir uns mit den drei Kernproblemen dieser Disziplin auseinandergesetzt. Diese umfassen die Quantifizierung von Metaboliten, deren Identifizierung sowie statistische Methoden zur Erkennung charakteristischer Metabolitenmuster und neuer Biomarker. Unser Hauptaugenmerk lag dabei auf dem Design robuster und umfassender Workflows, wobei die genannten Kernfragen nicht isoliert, sondern im Zusammenhang betrachtet wurden.

Wir entwickelten einen neuen Algorithmus, welcher eine robuste Detektierung und Quantifizierung von Metabolitensignalen (*Features*) in LC-MS-Daten ermöglicht. Dieser extrahiert chromatographische Profile mit hoher Sensitivität und sieht eine Auflösung von Isotopenmustern mittels eines neu entwickelten SVM-Modells vor. Unser Algorithmus wurde sowohl auf echten als auch simulierten LC-MS-Benchmarkdatensätzen validiert und zeigte im Vergleich zu bestehenden Lösungen hervorragende Ergebnisse. Zur Identifizierung von unbekannten Features entwarfen wir eine umfassende, integrative Strategie, welche sich möglichst viele Eigenschaften eines Features (z.B. Isotopenmuster, Retentionszeit) zu Nutze macht. Diese wurden dann zu Filterkriterien kombiniert, um eine verlässlichere Identifikation zu erzielen. Zu diesem Zweck entwickelten wir ein Softwaretool, mit dem eine hohe Anzahl an Features mittels ihrer Masse effizient gegen eine Metabolitendatenbank abfragt werden kann. Standardmäßig deckt es eine breite Palette an potentiellen Addukten ab, kann aber auch frei konfiguriert werden. Wir erweiterten unser Suchwerkzeug um einen Isotopen- und Retentionszeitfilter, mit dem Ziel, die Zahl der falsch positiven Treffer beträchtlich zu reduzieren. Der Nutzen von Modellen zur Vorhersage der Retentionszeit wurde zwar vereinzelt nachgewiesen, routinemäßig erstellt und in ID-Pipelines integriert — wie in unserem Fall — wurden sie jedoch nicht. Des Weiteren implementierten wir eine effiziente Methode zum Abgleich von MS/MS-Spektren gegen eine eigens zugeschnittene Spektrendatenbank. Unsere Methoden wurden auf Grundlage eines LC-MS/MS-Datensatzes validiert und erzielten in ihrem Zusammenspiel eine hervorragende Genauigkeit.

Das übergeordnete Ziel der "Computational Metabolomics" ist es, aus komplexen LC-MS-basierten Metabolomikdaten neue biologische Erkenntnisse zu gewinnen. Hierfür entwarfen wir umfassende Workflows zur Analyse von globalen Metabolomikdaten, die auf robusten statistischen Methoden aufbauten. Ihren Nutzen konnten wir zeigen, indem wir zwei biologisch relevante Fragestellungen beantworteten. Die erste zielte auf die Rolle des rs7903146 Polymorphismus im TCF7L2-Gen und dessen Auswirkung auf T2DM ab. Wir entdeckten mehr als 100 potentielle Biomarker, von denen einige auf für T2DM typische Störungen in der Biosynthese von Gallensäuren und Androgenen hinweisen. Unser Ansatz übertraf klassische statistische Methoden wie etwa PLS-DA, welche keinerlei signifikante Unterschiede zwischen den Kontroll- und Risikoallelgruppen aufzeigen konnte. Die zweite Fragestellung betraf das Auffinden von charakteristischen Mustern und entsprechenden Stoffwechselwegen, die während einer zweistündigen Belastungsphase und einer darauffolgenden dreistündigen Ruhephase beeinflusst wurden. Wir adaptierten einen bekannten Clusteralgorithmus, um eine Vielzahl von solchen Zeitreihen zu komprimieren, die aus nur wenigen und nicht gleichabständigen Zeitpunkten bestehen — ein häufig anzutreffender Fall bei Zeitreihenanalysen auf Metabolomikdaten. Unser neuartiger Clustering-Ansatz brachte 25 unterschiedliche Cluster hervor, die jeweils für die Belastungs- oder Erholungsphase repräsentativ waren. Die zwei auffälligsten Cluster legten im Rahmen einer "Pathway Enrichment" Analyse nahe, dass Aminosäure-, Fettsäure- und Katecholamin-Stoffwechselwege involviert waren. Veränderungen innerhalb dieser Stoffwechselwege im Zusammenhang mit körperlicher Betätigung und Erholung wurden in der Forschung bereits diskutiert. Beide Screening-Workflows führten zu biologisch aussagekräftigen Ergebnissen, mit Hilfe derer wir spezifischere Hypothesen aufstellen konnten. Dank dieser können wir nun unsere Ergebnisse zielgerichtet mit weiterführenden Experimenten validieren, während redundante Analysen vermieden und somit zeitliche und finanzielle Ressourcen eingespart werden können.

Contents

1 Introduction 1

2 Background 9

2.1 Metabolomics . . . 9

2.1.1 Liquid Chromatography Coupled to Mass Spectrometry . . . 11

2.1.2 Targeted versus Untargeted Metabolomics . . . 15

2.2 Computational Metabolomics . . . 16

2.2.1 Quantification . . . 16

2.2.2 Identification . . . 18

2.3 Statistical Methods in Metabolomics . . . 24

2.3.1 Data Pretreatment Prior to the Statistical Analysis . . . 27

2.3.2 Biomarker Detection with Rank Products . . . 28

3 Automated Label-Free Quantification of Metabolites from LC-MS Data 31

3.1 Introduction . . . 31

3.2 Algorithm . . . 33

3.2.1 Mass Trace Detection . . . 34

3.2.2 Feature Assembly . . . 37

3.3 Datasets . . . 42

3.3.1 Quantification Dataset . . . 42

3.3.2 Simulated Plant Metabolites Dataset . . . 44

3.4 Results . . . 46

3.4.1 Quantification Linearity and Reproducibility . . . 47

3.4.2 Recall, Precision, and Robustness of the Algorithm . . . 50

3.4.3 Comparison to Related Methods . . . 51

3.4.4 Availability . . . 52

3.5 Discussion 52

4 Metabolite Feature Identification 55

4.1 Introduction 55

4.2 Methods 57

4.2.1 RPLC-MS/MS Validation Dataset 57

4.2.2 Accurate Mass Search with Isotope Ratios Filtering 58

4.2.3 Spectral Search in Fragment Spectra Databases 61

4.2.4 Exploiting the AMS and MSM Tools in a Combined Approach 64

4.2.5 Prediction of Metabolite Retention Times on RPLC-MS Columns 66

4.3 Results 66

4.3.1 Individual Performance of AMS 68

4.3.2 Individual Performance of MSM 71

4.3.3 Validation of the Combined Metabolite ID Approach 72

4.3.4 Retention Time Prediction Models 72

4.4 Discussion 73

5 Analysis of Large-Scale Time-Resolved Metabolomics Data 77

5.1 Methods 79

5.1.1 Ethic Statement, Subjects, and Study Design 79

5.1.2 Metabolite Signal Extraction 79

5.1.3 Metabolite Identification 80

5.1.4 Cluster Analysis 80

5.1.5 Pathway Enrichment Analysis 82

5.1.6 UPLC-qTOF-MS Analysis 82

5.2 Results 82

5.3 Discussion 84

6 Untargeted Metabolomics Analysis for Metabolic Diseases 87

6.1 Introduction 87

6.2 Methods 88

6.2.1 Study Design 88

6.2.2 Bioinformatics Analysis 89

6.2.3 Untargeted Metabolomics Analysis 90

6.2.4 Clinical analyses 91

6.3 Results 91

6.4 Discussion 93

7 Conclusion 97

A Supplementary Material 101

A.1 File Archive 101

A.1.1 File Listing of Chapter 3 101

A.1.2 File Listing of Chapter 4 102
A.1.3 File Listing of Chapter 5 103
A.1.4 File Listing of Chapter 6 104
A.2 Figures 105
A.3 Tables 123

B Curriculum Vitae **141**

References **145**

List of Figures

1.1 The omics cascade of systems biology 2

1.2 Central problems in computational metabolomics 4

2.1 The electrospray ionization (ESI) method 13

2.2 Schematics of Thermo Fisher Scientific's Q Exactive Orbitrap platform 15

2.3 General computational quantification workflow 19

2.4 Outline of a typical metabolite ID workflow 21

2.5 Comparison of *cis*-7- and *cis*-9-hexadecenoate with respect to MS/MS fragmentation 25

2.6 Consensus matrix layout prior to the statistical analysis 29

3.1 General procedure of feature finding 34

3.2 Separation of chromatographic peaks within the same mass trace 37

3.3 Characteristics of a metabolite isotope pattern measured in LC-MS 38

3.4 Linear regression models approximating m/z spacings within metabolite isotope patterns 40

3.5 Relationship between Reserpine and Terfenadine concentrations and respective feature intensities 47

3.6 Numbers of reproducible features yielded by our algorithm and XCMS/CAMERA in the spike-in human plasma dataset 48

3.7 Intensity variation between features that were matched in at least half of the 33 spike-in human plasma measurements 49

4.1 Design of the search strategy as implemented by the AMS tool 58

4.2 The mapping relation R_{AMS} modeled by the AMS tool 60

4.3 General data extraction strategy of the `parseMassBank.py` script 62

4.4 E-value computation based on an exemplary HyperScoreModified distribution . 64

4.5 KNIME workflow for the computation of molecular descriptors 67

4.6 KNIME workflow for the training of an rt model and its application 67
4.7 Heat map visualization of AMS results 70
4.8 Impact of the RIA filtering on the number of candidate IDs and the percentage of annotated features 71
4.9 Retention time models employed as orthogonal filters for putative IDs 73
4.10 Densities of candidate ID numbers per feature before and after rt filtering 74

5.1 Clusters representative for metabolite perturbations in the exercise and recovery phases 83

6.1 Levels of androsterone/DHT sulfate and androstanediol glucuronide with respect to gender/group 93
6.2 Retention time models based on the TCF7L2 dataset 94

A.1 Relationship between propionyl-L-carnitine-d3 and nialamide concentrations and respective feature intensities 105
A.2 Relationship between sulfadimethoxine-d6 and hexadecanoyl-L-carnitine-d3 concentrations and respective feature intensities 106
A.3 Relationship between octadecanoyl-L-carnitine-d3 concentrations and feature intensities 107
A.4 Feature finding performance with respect to varying degrees of simulated detector noise 108
A.5 Feature finding performance comparison with respect to varying settings of the `mass_error_ppm` parameter 108
A.6 Feature finding performance with respect to varying settings of the `chrom_fwhm` parameter 109
A.7 Snapshot of the `INIFileEditor` window 110
A.8 UML diagram of the `AccurateMassSearchResult` class 111
A.9 Parsing of pseudo-molecular formulations in the AMS tool 112
A.10 Distribution of HMDB compound masses below 1,200 Da 112
A.11 Two-dimensional histogram of HMDB compound masses 113
A.12 Distribution of positive adduct types with respect to mass error tolerance $|\delta m_{\text{AMS}}|$114
A.13 Distribution of negative adduct types with mass error tolerance $|\delta m_{\text{AMS}}| \leq 5\,\text{ppm}$114
A.14 Clusters representative for the exercise phase (part 1) 115
A.15 Clusters representative for the exercise phase (part 2) 116
A.16 Clusters representative for the recovery phase (part 1) 117
A.17 Clusters representative for the recovery phase (part 2) 118
A.18 Profile plots of an up-regulated candidate detected in positive mode 119
A.19 Profile plots of a down-regulated candidate detected in positive mode 120
A.20 Profile plots of a down-regulated candidate detected in negative mode 121
A.21 Profile plots of a down-regulated candidate detected in negative mode 122

List of Tables

2.1 The MSI guidelines proposed for metabolite ID 20

3.1 Standard compounds employed in the spike-in experiments 43
3.2 Performance scores computed on the basis of the simulated plant metabolite dataset 51

4.1 Overall statistics of AMS results from the validation dataset 68

5.1 Significantly enriched metabolic pathways in the exercise and recovery stages . 84

A.1 Configuration of detection parameters for the MS measurements conducted in positive ESI mode 123
A.2 Parameter configuration of `MSSimulator` 123
A.3 Feature intensities for spiked-in compounds 124
A.4 Structure of the AMS tool's tabular HMDB mapping file 125
A.5 Example for an `mzTab` output file 125
A.6 Abridged content of the `KNA00121` record from MassBank 126
A.7 Validation dataset for our metabolite ID strategies 127
A.8 Training datasets for the RPLC rt prediction models 131
A.9 Characteristics of the individuals participating in the one-legged knee extensor experiments 131
A.10 Subject characteristics of our TCF7L2 cohort 132
A.11 Up-regulated metabolites detected in the TCF7L2 positive mode dataset 133
A.12 Putative AMS IDs of up-regulated candidates from positive mode data 134
A.13 Down-regulated metabolites detected in the TCF7L2 positive mode dataset . . . 135
A.14 Putative AMS IDs of down-regulated candidates from positive mode data 136
A.15 Up-regulated metabolites detected in the TCF7L2 negative mode dataset 137

A.16 Putative AMS IDs of up-regulated candidates from the TCF7L2 negative mode data 138
A.17 Down-regulated metabolites detected in the TCF7L2 negative mode dataset . . . 139
A.18 Putative AMS IDs of down-regulated candidates from the TCF7L2 negative mode data . . . 140

Acronyms

2D two-dimensional. 32, 70

3D three-dimensional. 32

AGC automatic gain control. 57

AMS `AccurateMassSearch`. 6, 48, 55–61, 63–66, 68–70, 72–76, 78, 80, 83, 85, 90, 94, 98–100, 111, 112, 114, 119–122, 125, 127–131, 134, 136, 138, 140

ANOVA analysis of variance. 7, 25

ATP adenosine triphosphate. 1

BMI body mass index. 89, 131, 132

BP blood pressure. 131

CA cholic acid. 95, 131

CDCA chenodeoxycholic acid. 92, 94, 95, 140

CDK chemistry development kit. 67

CE-MS capillary electrophoresis-mass spectrometry. 10, 23, 33

CID collision-induced dissociation. 24, 25

CRP C-reactive protein. 89

CV coefficient of variation. 28, 49, 89

DB database. 6, 20, 24, 58, 65, 71, 72, 76, 99, 100

DCA deoxycholic acid. 92, 140

DDA data-dependent acquisition. 73, 99

DG diglyceride. 138

DHT dihydrotestosterone. 92, 93, 95, 121

DIMS direct infusion mass spectrometry. 10

DNA deoxyribonucleic acid. 1, 2

DP dynamic programming. 20

EDTA ethylenediaminetetraacetic acid. 43, 90

EI electron ionization. 12

ESI electrospray ionization. 12, 22, 31, 32, 44, 75, 76, 82, 90, 91, 100, 123

FC fold change. 26, 28–30

FDR false discovery rate. 25, 30, 89

FFA free fatty acid. IV, 91, 92, 95, 98, 131

FFM `FeatureFinderMetabo`. 5, 6, 33, 49, 51, 52, 55, 57, 99, 100

FTICR Fourier transform ion cyclotron resonance. 10, 11, 22, 75

FWHM full width at half maximum. 40, 42

GC gas chromatography. 23, 78

GC-MS gas chromatography-mass spectrometry. 10–12, 23

GCA glycocholic acid. 131

GCDCA glycochenodeoxycholic acid. 131

GWA genome-wide association. 88

HCD higher-energy collisional dissociation. 15, 24

HILIC hydrophilic interaction chromatography. 23

HMDB Human Metabolome Database. 10, 23, 41, 48, 55, 58, 59, 69, 75, 80, 82, 85, 91–94, 112, 113, 125, 134

HOMA homeostasis assessment model. 91, 95, 132

HPLC high performance liquid chromatography. 33, 43, 57

ICR ion cyclotron resonance. 12

ID identification. III, IV, 5, 6, 9, 10, 12, 16, 18, 20–24, 48, 55–58, 61, 63, 65–76, 80–83, 85, 90–96, 98–100, 119–122, 127–131, 133–140

iIGT isolated impaired glucose tolerance. 89

InChI international chemical identifier. 59, 66, 67

KEGG Kyoto Encyclopedia of Genes and Genomes. 85

KNIME Konstanz information miner. 52, 66, 67

LC liquid chromatography. 11, 12, 23, 24, 75, 76, 91

LC-MS liquid chromatography-mass spectrometry. III–VI, 3–8, 10–12, 17, 18, 23, 24, 31–35, 38, 43, 45, 47, 52, 53, 62, 73, 76–79, 85, 90, 93, 97, 98, 100

limma linear models for microarray data. 26

LOD level of detection. 27, 28, 99

LOO leave-one-out. 72

LOWESS locally weighted scatterplot smoothing. 36

LPC lysophosphatidylcholine. 131, 134, 136

LPE lysophosphatidylethanolamine. 131, 134, 138

LTQ linear trap quadrupole. 14

m/z mass-to-charge ratio. 3, 5, 6, 12, 17, 24, 32–35, 37–46, 50, 57, 58, 61, 63, 65, 66, 68, 69, 79, 80, 82, 89, 90, 92, 98, 113, 123, 133–140

MCC Matthews correlation coefficient. 41

MCP Money Changing Problem. 20

MLR multiple linear regression. 23

MS mass spectrometry. III, IV, VI, 3, 5, 9–13, 16, 17, 19, 20, 22, 23, 29, 33, 35, 36, 42–45, 52, 56, 57, 60, 65, 66, 69, 75, 76, 81, 82, 89, 90, 95, 100, 123

MS/MS tandem mass spectrometry. IV, VI, XI, 3, 5, 6, 14, 15, 18, 20, 24, 25, 56, 57, 64–66, 71, 73, 75, 76, 85, 95, 99, 100, 126

MSI Metabolomics Standards Initiative. 5, 6, 18, 20, 99

MSM `MetaboliteSpectralMatcher`. 6, 56–58, 61, 63–66, 68, 71–76, 99, 100, 127–130

NCE normalized collision energy. 57

NEFA non-esterified fatty acids. 91

NIST National Institute of Standards and Technology. 76

NMR nuclear magnetic resonance. 10, 11, 18, 78

OGTT oral glucose tolerance test. 91

OPLS orthogonal partial least squares regression. 26

OPLS-DA orthogonal projections to latent structures discriminant analysis. 7, 26

PaDEL Pharmaceutical Data Exploration Laboratory. 66, 67

PC phosphatidylcholine. 93, 134, 136

PCA principal component analysis. 26, 27, 77, 95, 98

PE phosphatidylethanolamine. 134, 138

PFP percentage of false positives. 30, 89, 133, 135, 137, 139

PIC pure ion chromatogram. 32

PLS partial least squares regression. 26

PLS-DA partial least squares discriminant analysis. IV, VI, 7, 26, 77, 95, 98

QC quality control. 28, 29, 89, 91

QT quality threshold. 7, 78, 81–84, 98

QTOF quadrupole time-of-flight. 12, 22, 44, 82

RF radio frequency. 14, 15

RIA relative isotopic abundance. III, XII, 5, 6, 22, 41, 55, 56, 60, 61, 65, 68, 69, 71, 72, 74, 75, 90, 94, 98, 99

RMSE root mean squared error. 22, 41, 61, 69, 75, 93

RNA ribonucleic acid. 1, 2

RP rank product. 28–30

RPLC reversed-phase chromatography. 6, 11, 23, 56, 57, 64–66, 73, 75, 76, 81, 99, 100, 131

rt retention time. III, 3, 5, 6, 10, 12, 16, 17, 21, 23, 33–36, 38, 41–43, 45, 46, 52, 56, 65–69, 72–76, 78–80, 83, 85, 89, 90, 93, 94, 99, 131, 133–140

SAM significance analysis of microarrays. 26

SD standard deviation. 132

SDF structure data format. 67

SMILES simplified molecular input line entry specification. 59

SMPDB Small Molecule Pathway Database. 78, 80, 82, 84, 85, 98

SNP single nucleotide polymorphism. 7, 88, 91

STS short time-series. 7, 78, 81, 84, 98

SVM support vector machine. III, V, 5, 33, 41, 42, 99

T1DM type 1 diabetes mellitus. 89

T2DM type 2 diabetes mellitus. IV, VI, 2, 7, 28, 87, 88, 91, 92, 94, 95, 97, 100

TCA tricarboxylic acid. 2, 57, 72, 127, 128

TCDCA taurochenodeoxycholic acid. 140

TCF7L2 transcription factor 7-like 2. IV, VI, XII, 7, 88, 94, 95, 97, 100, 132, 133, 135, 137–140

TDCA taurodeoxycholic acid. 140

TIC total ion chromatogram. 17

TOF time-of-flight. 12, 41

TOPP the OpenMS proteomics pipeline. 33, 52

TUEF Tuebingen family. 88

UML unified modeling language. 111

UPLC ultra performance liquid chromatography. 11, 43, 44, 82, 90

XIC extracted ion chromatogram. 5, 17, 90, 92

XML extensible markup language. 42, 52, 59

CHAPTER 1

Introduction

Our genetic code is the blueprint of our lives. It is thus not surprising that research has undertaken huge efforts over the last century to understand its role. While we have made considerable progress in this field, in particular, through deciphering the sequence of the human genome at the outset of this century, we also came to realize that "knowing the code" alone does not suffice to explain our genes' implications on our lives. From the central dogma of molecular biology, we know that there are several layers between an organism's genetic background, the *genotype*, and the actual occurrence of a biological trait, the *phenotype* (Figure 1.1). Each layer comes with its own array of regulation mechanisms and level of complexity. Genes are first copied from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) during *transcription*. These transcripts serve as templates for the biosynthesis of new proteins during the *translation* process. Proteins can fulfill manifold functions in a biological system, e.g., as building blocks, molecule transporters, cell signal carriers, or catalyzing biochemical reactions as *enzymes*. With these roles, proteins have a more direct and evident impact on biological processes than the genetic level. Hence, it is not surprising that they are studied extensively in order to explain specific phenotypes. However, the protein level is still not the last line of players that immediately influences a phenotype. A more complete view of a phenotype's biological background can be obtained if we go one step further and focus on the small molecules that are subject to enzymatic reactions, the *metabolites*.

Metabolites owe their name to the Greek word *metabolē* (*to change*). Their naming reflects the fact that they are constantly subject to breakdown and synthesis reactions that constitute a biological system's *metabolism*. The metabolism can be roughly divided into *catabolism*, the breakdown and conversion of energy-rich molecules (e.g., from food) into the readily available energy currency adenosine triphosphate (ATP), and *anabolism*, the synthesis of polymeric

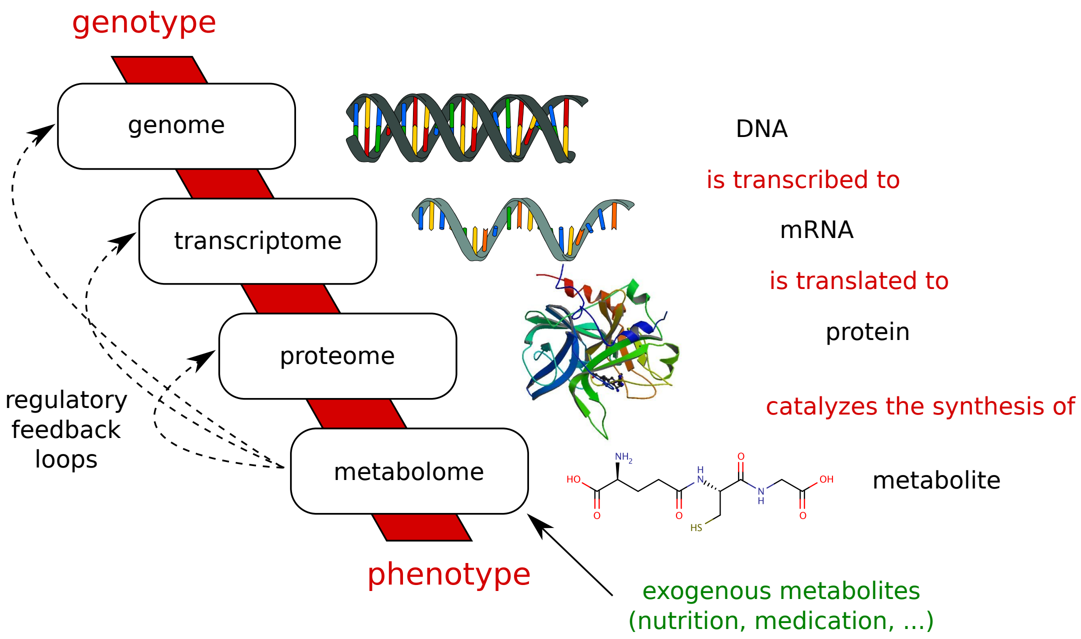


Figure 1.1: The omics cascade of systems biology. The connection between the domains closely resembles the central dogma of molecular biology. We adapted the DNA and RNA graphics from [1], the protein image from [2], and the metabolite image from [3].

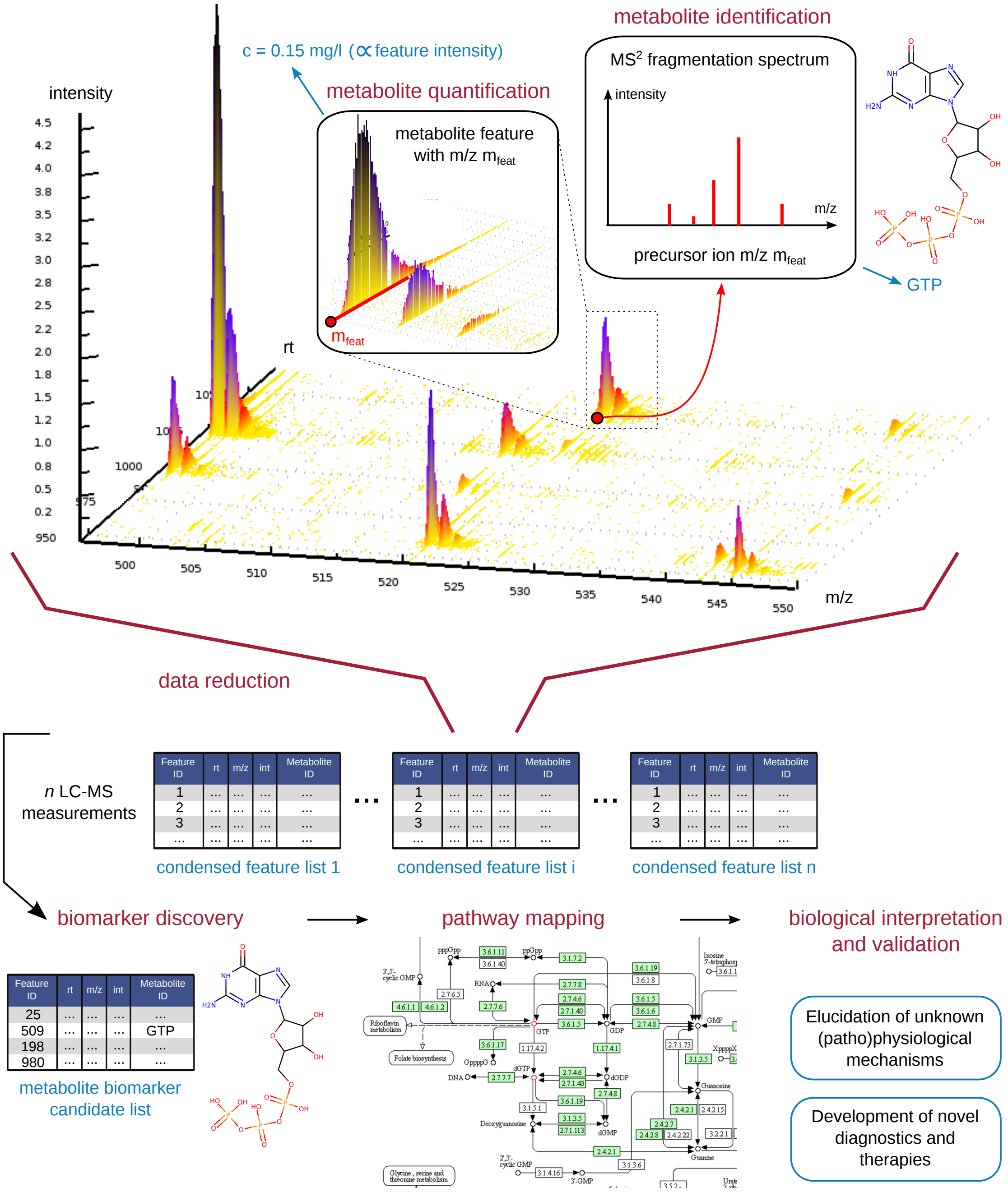
molecules (proteins, lipids, carbohydrates) in order to provide and maintain cell structure and function. Early biochemical studies were focused on very specific aspects of metabolism, for example, to understand the utilization of glucose in the fermentation process. Long before the first genes involved were sequenced, the *glycolysis* pathway and the tricarboxylic acid (TCA) cycle were elucidated experimentally in the 1930s and 1940s, two highly conserved mechanisms that are central to energy metabolism in most biological organisms. Since then, numerous other functionally coherent subcomponents in metabolism were formulated as *metabolic pathways* that were mapped meticulously, most prominently, in the *biochemical pathways* wall charts [4]. The growing understanding of the metabolism enabled us to explain various phenotypes, particularly, those induced by disease, diet, or environmental changes. For instance, the hereditary disease *phenylketonuria* was attributed to a defect in the enzyme *phenylalanine hydroxylase* that caused the accumulation of phenylalanine and with that neurological damage in newborns if not treated by a special diet. While the phenotype could be pinned down on one specific metabolic pathway in this case, there are multifactorial diseases such as T2DM that affect multiple pathways at the same time and thus are much more complex to study with the classical mechanistic experiments.

Instead of focusing only on a few metabolites that could cause a disease, a more holistic approach had to be followed in order to detect multiple factors or complex *patterns*. Such an approach had to provide simultaneous measurement and analysis of the entirety of metabolites present in a biological system, the *metabolome*. Analogously to how *genomics* was coined to refer to the analysis of the *genome* [5] and later was adopted for the *transcriptome* and *proteome*, the comprehensive and systematic analysis of the metabolome became known as *metabolomics*.

In the early years of metabolomics studies, the major challenge was the development of a robust analytical platform that could deal with the high complexity and chemical diversity of metabolomes [6]. For instance, the chemical analysis in proteomics experiments could focus on peptides exclusively while metabolomics had to capture various compound classes with different physicochemical properties such as carbohydrates, lipids, or bile acids. Furthermore, the analytic method of choice had to be sensitive enough to capture compounds in extremely low concentrations (e.g., signaling molecules). Both the high complexity and the wide range of metabolite concentrations were common scenarios when complex biological samples such as blood plasma had to be analyzed. An analytical platform meeting these requirements was established with LC-MS.

In an LC-MS experiment, the metabolites of a complex sample are first separated by rt with respect to their physicochemical properties and then their ionized counterparts are resolved by mass-to-charge ratio (m/z). The mass spectrometer quantifies the numbers of ions with the same m/z and records their *intensities* in a *mass spectrum*. In MS/MS, specific ions can be guided into a collision cell where they are fragmented, and the resulting pattern is recorded as a *fragmentation spectrum* for identification purposes. The complete set of mass spectra collected over rt yields complex two-dimensional datasets (or LC-MS maps) that can contain tens of thousands of signals (Figure 1.2). Recent MS technologies such as the Orbitrap mass spectrometer can deliver very high resolutions with short cycle times per spectrum and thus easily produce several gigabytes per LC-MS measurement. Since the manual interpretation of such complex datasets is not timely feasible, efficient computer algorithms for automated data processing and analysis have become indispensable and thus gave rise to the field of *computational metabolomics*.

Computational metabolomics aims at the extraction of biologically relevant information from the LC-MS data by addressing three main problems (Figure 1.2). First, the signals observed in the LC-MS map and induced by metabolites (so-called *features*) must be extracted and their concentrations quantified in a robust and reproducible manner. Second, the identities behind these features must be revealed and validated before any biological interpretation can take place. Third, the detection of differential metabolite patterns (e.g., induced by a disease phenotype) and their potential biological interpretation must be assisted with sound statistical methods. The latter usually aims at the discovery of metabolite biomarkers and mapping them to metabolic pathways that explain the underlying mechanisms of a (patho)physiological condition and assist with the development of novel disease diagnostics [7]. In recent years, there has been active research addressing these central problems by many methods and algorithms. Although considerable progress has been made in each of the areas, the presented solutions were usually limited to the specific problem and could not be readily integrated with existing computational pipelines. Common limitations were incompatibilities due to different implementations or operating systems and missing or inconsistent exchange formats between software tools. Often, solutions were implemented as standalone monolithic programs that were not modular enough to be part of more complex automated workflows. In this work, we address the central problems


metabolite identification
c = 0.15 mg/l (∝ feature intensity)
metabolite quantification
MS^2 fragmentation spectrum
intensity
m/z
precursor ion m/z m_{feat}
GTP
intensity
metabolite feature with m/z m_{feat}
m_{feat}
rt
m/z
4.5
4.2
4.0
3.8
3.5
3.2
3.0
2.8
2.5
2.2
2.0
1.8
1.5
1.2
1.0
0.8
0.5
0.2
950
975
1000
500
505
510
515
520
525
530
535
540
545
550
data reduction
n LC-MS measurements
condensed feature list 1
condensed feature list i
condensed feature list n
biomarker discovery
pathway mapping
biological interpretation and validation
metabolite biomarker candidate list
Elucidation of unknown (patho)physiological mechanisms
Development of novel diagnostics and therapies

Feature ID	rt	m/z	int	Metabolite ID
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

Feature ID	rt	m/z	int	Metabolite ID
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

Feature ID	rt	m/z	int	Metabolite ID
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

Feature ID	rt	m/z	int	Metabolite ID
25	...	...	...	...
509	...	...	...	GTP
198	...	...	...	...
980	...	...	...	...

Figure 1.2: Central problems in computational metabolomics. LC-MS yields complex data with thousands of metabolite signals that first must be processed computationally. Metabolite signals are detected and quantified proportionally to their observed intensity. The extraction of these metabolite *features* from noisy data usually results in a massive *data reduction*. If the same feature mass m_{feat} was fragmented, its fragment spectrum can be matched against a spectral database to identify the measured metabolite. Statistical methods are applied on the condensed feature lists of multiple LC-MS measurements to detect potential biomarkers. Mapping these back to metabolic pathways could explain the underlying biological mechanisms and result in a better understanding and predictability of a disease phenotype.

of computational metabolomics with a strong focus on establishing a coherent and consistent workflow that still allows for high customization due to its modularity. In the following, we elaborate on each of the key problems and present our contributions, respectively.

LC-MS measurements of biological samples produce complex datasets not only with large numbers of metabolite signals but also with signals caused by *chemical noise* (e.g., due to unexpectedly fragmented compounds), *technical noise*, and contaminants. In order to quantify the concentrations of metabolites, the corresponding features must be extracted robustly before the background of noise. To this end, a typical *feature detection* algorithm usually attempts to capture the bell-shaped chromatographic profiles (Figure 1.2), each induced by an analyte ion with a specific m/z. Since MS also resolves the naturally occurring isotopes of analytes, feature detection algorithms should also be capable of resolving isotope patterns such that the redundancy introduced by satellite ^{13}C isotopes is avoided. In this regard, conflicts of overlapping isotope patterns introduced by coeluting analytes must be dealt with. Several methods and algorithms have been proposed to cope with the feature detection problem, either partially or completely (Section 3.1). An R package [8] widely used in the metabolomics community is XCMS [9]. The XCMS algorithm focuses on the extraction of chromatographic profiles or extracted ion chromatograms (XICs) only and does not provide any means to assemble isotope patterns per se. The latter functionality is added by CAMERA, another R package building on the XCMS results [10]. Chromatographic profiles are detected with a second order matched Gaussian filter, that is, the profile shapes are assumed to be Gaussian. This assumption is often violated by asymmetrical profile shapes (e.g., heavy-tailed profiles). More importantly, the matched filter may not be sensitive enough to detect the low-intensity signals of metabolites in extremely low concentrations, a common scenario for biological samples. In Chapter 3, we present our novel feature detection algorithm FeatureFinderMetabo (FFM) for the automated and label-free quantification of metabolites. In contrast to XCMS, our algorithm makes no assumptions of the chromatographic profile shape and thus can detect low-intensity metabolites with high sensitivity. Furthermore, it tackles the problem of deisotoping features with a novel SVM classifier that is specifically trained to recognize valid isotope abundances of organic small molecules. We validated the algorithm with human plasma LC-MS data and found excellent linearity between the spike-in compound concentrations and the quantified feature intensities. In benchmarks based on simulated LC-MS data, our algorithm clearly outperforms the XCMS and CAMERA combination with respect to recall and precision. The modular implementation of FFM under the OpenMS framework [11] facilitates its integration into computational metabolomics pipelines (Section 2.2.1). Its ease of use (e.g., intuitive parameter configuration) and efficient run-times allow the researcher to focus more on the downstream analysis and biological interpretation of the complex LC-MS datasets.

The greatest challenge in computational metabolomics is the reliable identification of metabolites from LC-MS data. According to the guidelines proposed by the Metabolomics Standards Initiative (MSI), a metabolite feature's ID is considered as *confident* if at least two orthogonal properties (e.g., m/z, RIA, and rt) and its MS/MS fragmentation pattern can be matched to the corresponding authentic standard [12]. While several computational methods and algorithms to provide metabolite ID based on these criteria have been proposed (Section 2.2.2), the actual problem lies in the lack of comprehensive, consistent, and readily available resources of orthogonal metabolite data based on LC-MS. Both the unknown compounds and the authentic standards must be measured under identical conditions (i.e., experimental setup, measurement protocol). Consequently, a representative database (DB) of authentic standards must be built first before any comprehensive metabolite ID can take place. Such an endeavor is usually very expensive and time-consuming due to high costs of standard compounds and numerous LC-MS and MS/MS measurements. Thus, it is only feasible for a small subset of metabolites (e.g., for the targeted analysis of lipids). Since such compound DBs are built on very specific analytical setups, their transferability to other LC-MS platforms is rather limited. For example, the rt information and fragmentation spectra of metabolites are hardly comparable if they were measured with different LC-MS columns (high variability between manufacturers and even between batches) and fragmentation setups (distinct collision cells and energies), respectively. The availability of public databases is crucial particularly for untargeted metabolomics experiments (Section 2.1.2). Here, the putative identification of metabolite patterns is essential to generate novel hypotheses even if the IDs are considered incomplete or less confident according to the MSI guidelines. Follow-up experiments can then address these hypotheses in a targeted fashion, reducing time and cost for the validation of metabolite IDs. In this work, we approached these problems and developed a computational metabolite ID workflow that leverages and combines all sources of orthogonal information that is extractable from untargeted LC-MS data to boost both the number of IDs and their confidence (Chapter 4). It readily integrates with our quantification pipeline and provides the automated annotation of metabolite features detected by our FFM algorithm. We implemented `AccurateMassSearch` (AMS), an efficient and flexible software tool for the rapid detection of a wide range of potential metabolite adducts based on a feature's m/z (Section 4.2.2). With the integrated RIA filter, AMS significantly reduces the number of incorrectly assigned metabolite adducts. This number can be further decreased by the application of an rt filter. We developed a computational method that automatically generates an rt prediction model based on the computed molecular properties of candidate IDs that were confirmed by two or more orthogonal criteria. By comparing the observed and predicted rt, candidate IDs with unrealistic rt differences are discarded. Such an rt filter can be a powerful criterion since it combines chemical knowledge with the otherwise mass-centric AMS approach. If MS/MS spectra are present, they can be matched against the MassBank spectral repository [13] with our `MetaboliteSpectralMatcher` (MSM) tool. MSM was designed to compute the spectral matching of thousands of fragmentation spectra efficiently and to provide reliable matching and confidence scores. Both the AMS and MSM results are reported in the mzTab standard [14] and are merged in a final step to achieve the consensus IDs. Based on a reversed-phase chromatography (RPLC)-MS/MS validation dataset, we could show that the identification rates of our combined approach were considerably better than the tools' individual performances.

The ultimate goal of a complete computational metabolomics workflow is to turn the information gained from the metabolite quantification and identification into novel biological knowledge. Due to the complexity and size of such datasets, sound statistical methods are necessary to extract metabolite patterns that are predictive of the phenotype of interest. In general, statistical methods must be sensitive enough to detect subtle fold changes of metabolite levels between treatment groups and must deal with strong intra-group variations [15]. Proper treatment of data before statistical analysis is equally important. It comprises handling of missing values, inter-sample intensity normalization, and removal of noisy or spurious observations. Univariate statistics such as the t-test, the nonparametric Wilcoxon test, or the analysis of variance (ANOVA) are inadequate to detect significant metabolite patterns in high-dimensional datasets if no measures to deal with the problem of *multiple hypothesis testing* are taken [16]. Furthermore, statistical tests such as the t-test are not readily applicable to time-course data since the assumption of independence between time points is usually not reasonable. Several multivariate methods such as the prominent PLS-DA and its variants were studied extensively for metabolomics biomarker discovery (Section 2.3). PLS-DA has become a standard tool to rapidly visualize the differences between treatment groups and to extract the metabolite patterns responsible for the separation. However, PLS-DA models must be validated carefully to avoid overfitting. Depending on the diagnostic statistics employed for the model validation, the PLS-DA computation may prefer models with low complexity instead of building on metabolites with very small inter-group differences [17]. As a part of this work, we developed a robust biomarker discovery pipeline that is specifically geared towards the detection of low-concentration metabolite changes and easily integrates with our quantification and identification workflows (Section 6.2.2). It provides means to perform inter-sample normalization and to reduce data complexity by filtering out observations with high variation and low signal-to-noise ratio [15]. The statistical analysis is based on the nonparametric *rank product* test [18] which we adapted for the robust detection of significant fold changes in metabolomics data. With our discovery pipeline, we studied the relevance of the TCF7L2 rs7903146 single nucleotide polymorphism (SNP) for T2DM based on an LC-MS dataset of 30 individuals (15 controls versus 15 SNP carriers) (Chapter 6). In contrast to former studies where no significant differences between the groups were detected with the application of orthogonal projections to latent structures discriminant analysis (OPLS-DA) [19], we could extract 108 potential biomarkers with our pipeline. From these, we could find strong indications for the impairment of bile acid synthesis and differing testosterone levels in the TCF7L2 group, both known symptoms of T2DM [20, 21, 22]. In order to address the necessity to analyze time-resolved metabolomics data, we developed a novel computational approach that condenses thousands of time profiles into a few kinetic patterns. Our pipeline is based on an adapted version of the quality threshold (QT) clustering algorithm [23] for which we implemented the short time-series (STS) distance function [24]. This adapted version was particularly strong in clustering short time series with time points not being equidistant, a common scenario in metabolomics time-course studies. We validated our approach with LC-MS data that was collected at several time points during a single bout of exercise in male subjects. Based on this dataset, we performed a pathway enrichment analysis of the most interesting clusters (kinetic patterns with peaks during the exercise or recovery phase of the study) and identified several metabolic pathways. We assume that these pathways might play an important role during the exercise and recovery phases and thus we would suggest these as worthwhile subjects in a targeted metabolomics analysis.

In summary, this thesis makes valuable contributions to computational metabolomics by establishing a comprehensive workflow for the analysis of untargeted LC-MS data that addresses the central problems of metabolite quantification, identification, and biomarker discovery consistently. Instead of spending time on the details of the individual processing steps (e.g., due to cumbersome configuration or missing interfaces), metabolomics researchers can now focus more on the biological interpretation of their data and generate novel working hypotheses more quickly, hereby reducing overall time and costs of metabolomics studies.

CHAPTER 2

Background

This chapter introduces the main concepts of this work as much as the current state of the art. First, we give an overview of metabolomics and mass spectrometry (MS), its most common analytical platform. This is followed by a summary of current computational methods that were developed for the quantification and identification (ID) of metabolite signals in metabolomics datasets. We conclude the chapter with the presentation of a robust statistical analysis pipeline for the discovery of novel metabolite biomarkers.

2.1 Metabolomics

During the initial studies on the human genome mapping, researchers established the term *genomics* to refer to the systematic analysis of *all genes* [5]. Consequently, the comprehensive studies of the *transcriptome* and *proteome* in molecular biology were named in the same fashion, i.e., *transcriptomics* and *proteomics*. In close resemblance to the central dogma of molecular biology, the term "omics cascade" was coined to point out the relationship between the various omics levels [6] (Figure 1.1). Of particular interest to this work is *metabolomics*, the systematic study of all metabolites (the *metabolome*) found in a biological system. The link between the metabolomics and the other omics levels is given by the fact that metabolites are synthesized by metabolic reactions that are catalyzed and regulated by enzymes from the proteome. These enzymes, in turn, are tightly controlled by the genomic and transcriptomic levels (i.e., via gene expression). Changes on the metabolome level (e.g., increase/decrease of metabolite concentrations) usually show immediate effects on the organism's phenotype, whereas changes on the upstream levels (e.g., distinct gene expression profiles) may not directly result in phenotypic alterations. For this reason, metabolomics has become an indispensable tool to study phenotypic changes induced by disease, environmental influences, diet, or physical exercise.

While the analytical methods to study the genome, transcriptome, and proteome are well-established, the comprehensive analysis of an organism's metabolome with a single standardized analytical platform still remains a challenge. This is mainly due to the high chemical diversity of metabolomes. In contrast to proteomics, where the analytical method can be optimized to capture peptides only, metabolomics experiments must deal with a broad range of different compound classes (e.g., sugars, amino acids, or lipids) [25]. This is aggravated by the sheer number of different metabolites that can occur in complex samples such as blood plasma. In 2009, the Human Metabolome Database (HMDB) reported about 6,800 experimentally validated metabolites, many of which could be detected in human biofluids [26]. A more recent release of the HMDB database from the year 2013 that additionally integrated metabolites from exogenous sources (e.g, food, drugs) increased this number to more than 40,000 compounds [27]. More than 200,000 metabolites were estimated to exist in the plant kingdom [28]. The highly dynamic nature of the human metabolism facilitates the quick adaptation to different situations and external stimuli (e.g., physical activity, fasting, food intake, or medication) but also renders accurate and reliable measurements of its current state very difficult. The analytical method of choice must offer a broad dynamic range in order to capture the wide spectrum of metabolite concentrations and, in particular, must be sensitive enough to detect metabolites in extremely low concentrations (e.g., signaling molecules).

Early metabolomics studies made extensive use of nuclear magnetic resonance (NMR) technology as primary analytical platform [29]. Biological samples need no special preparation prior to the analysis, and thus the chance of unwanted contaminations or chemical reactions is low. By interpreting the NMR spectra, hundreds of metabolites can be quantified and identified in an automated fashion. However, the more complex a biological sample, the more difficult the interpretation of NMR spectra becomes. Furthermore, the NMR platform may fail to detect metabolites in very low concentrations, a scenario frequently encountered when working with biological samples. In order to complement the information missed by NMR and to facilitate a more comprehensive analysis of a biological sample, analytical methods based on MS were studied as an alternative. MS offers a higher sensitivity than NMR and is usually coupled to orthogonal separation techniques such as chromatography or electrophoresis. The upstream separation process results in time-resolved spectral information with lower complexity that is easier to interpret than a single spectrum containing the entirety of measured signals. The analytes' rts are characteristic for their physicochemical properties and may be exploited for compound ID (Section 4.2.5). Widely used MS-based analytical platforms are coupled either to gas chromatography (GC-MS), liquid chromatography (LC-MS), or capillary electrophoresis (CE-MS) [30]. In case of ultra-high resolutions as delivered by Orbitrap or Fourier transform ion cyclotron resonance (FTICR) MS, complex samples can be introduced directly into the mass spectrometer (direct infusion mass spectrometry (DIMS)) and analyzed on the basis of a single spectrum. Although these platforms offer higher sensitivity than NMR, each has its own limitations and thus does not cover the whole range of metabolites present in a complex biological sample [6]. For example, the GC-MS platform is limited to volatile compounds or those that can be modified accordingly by derivatization (e.g., by silylation [31]). In this work, we focus exclusively on LC-MS-based methods.

2.1.1 Liquid Chromatography Coupled to Mass Spectrometry

Technological advances in the field of high-resolution MS brought great benefits for metabolomics studies of complex biological samples [32]. High resolutions in the seven-figure range and mass accuracies below 1 ppm as provided by FTICR mass spectrometers facilitated the interpretation of both full-scan and tandem MS spectra by reducing the number of putative empirical formulas encountered with the annotation of spectral peaks. However, high operational and maintenance costs of such mass spectrometers limited their applicability for metabolomics studies. The FTICR mass analyzer employs a strong magnetic field (3–15 T) provided by superconducting magnets in order to trap ions in an orbital trajectory [33]. With the introduction of the Orbitrap mass analyzers, the high maintenance costs of cooling superconducting magnets could be circumvented by replacing the magnetic by an electric field [34]. Due to this cost-effective solution, Orbitrap-based mass spectrometers have become affordable to a wider audience and thus are now the analytical method of choice for many metabolomics studies. In the following, we present the principles of a liquid chromatography (LC)-coupled Orbitrap MS platform.

Reversed-Phase Liquid Chromatography

Prior to the actual MS measurement, biological samples are usually subjected to an LC separation step. The direct injection of complex samples into the mass spectrometer's ionization interface bears the risk of ion suppression effects [35] and often leads to overloaded spectra that are difficult to interpret. Instead, the sample complexity is resolved over time such that the mass spectrometer is confronted gradually with less complex fractions. In RPLC, the sample is mixed with an aqueous organic solvent (*mobile phase*) and pumped through a chromatographic column filled with hydrophobic material (*stationary phase*) [36]. Most often, a C18 column is employed as a stationary phase which contains silica with hydrophobic octadecanoyl ($C_{18}H_{37}$) chains on their surface. During the chromatographic run, the mobile phase is changed gradually in its water and organic solvent composition according to the *elution gradient*. By linearly increasing the organic solvent proportion over time, hydrophobic compounds strongly binding to the stationary phase will be carried away eventually. Another key factor of chromatographic separation is the *flow rate* of the mobile phase and the respective pumping pressure. Recent LC systems allow us to apply extremely high pressures yielding high flow rates on ultra performance liquid chromatography (UPLC) columns with minimal particle sizes. The advantage of these developments is reduced separation time while maintaining a high separation efficiency. In metabolomics studies, elution gradients in a range of 20 to 30 min are quite common. We refer to Section 6.2.3 for a detailed example of such a separation protocol. Each compound interacts specifically with both phases and thus travels at a distinct velocity. The time from a compound's injection to its appearance at the column's other end is called *retention time* (rt). Since the rt is characteristic of a compound's physicochemical properties and molecular structure, it provides valuable information to assist the ID of metabolites (Section 4.2.5).

Electrospray Ionization

The LC column is connected to the inlet of the mass spectrometer's ion source. Most of the incoming chemical compounds do not have an intrinsic charge and would not be detectable by MS without prior ionization. There are several methods that can be roughly classified by *hard* and *soft ionization*. As a classical hard ionization method, electron ionization (EI) generates ions from gas-phase molecules by pounding them with high-energy electrons. This usually results in the fragmentation of the original molecules, an issue that must be addressed when interpreting their spectra. EI is the method of choice for GC-MS experiments. Soft ionization methods on the contrary attempt to keep the molecule structure intact. One such method that proved itself invaluable when working with biomolecules in a LC-MS setting is electrospray ionization (ESI) [37]. Here, the sample liquid is sprayed through a thin capillary at atmospheric pressure to form an aerosol of charged droplets. When a high voltage is applied between the capillary and the mass spectrometer's inlet, a *Taylor cone* forms at the capillary's tip with charges aligning on the surface. Subsequently, charged droplets are sprayed from this cone into the ESI chamber. Heating can be applied to accelerate the solvent evaporation from these droplets; the droplets shrink until they burst into finer droplets due to charge-charge repulsions (*Coulomb fission*). The relationship between a droplet's size and its maximum capacity of charges is described by the *Rayleigh limit* [38]. Finally, when the solvent is completely evaporated, analyte ions remain that are directed into the mass spectrometer. ESI produces pseudo-molecular ions, that is, adducts that are formed by the interaction of uncharged polar molecules and small ions (e.g., proton or sodium ions) (Section 2.2.2).

Orbitrap Mass Analyzer

Aside from the ionization source and a mass detector, the central component of a mass spectrometer is the *mass analyzer* that provides us with an analyte's m/z. In the history of MS, several types of mass analyzers were developed, e.g., *magnetic sector field*, *quadrupole*, *time-of-flight (TOF)*, *ion trap*, *ion cyclotron resonance (ICR)*, or hybrids hereof such as the *quadrupole time-of-flight (QTOF)* analyzer [39]. For example, the quadrupole comprises four parallel cylindrical electrodes with oscillating electrostatic fields. By varying the electrode voltages, the quadrupole acts as a *mass filter* since only ions with a specific m/z or within a specific m/z range will have a stable trajectory to reach the quadrupole's other end without hitting the rods. The most recent

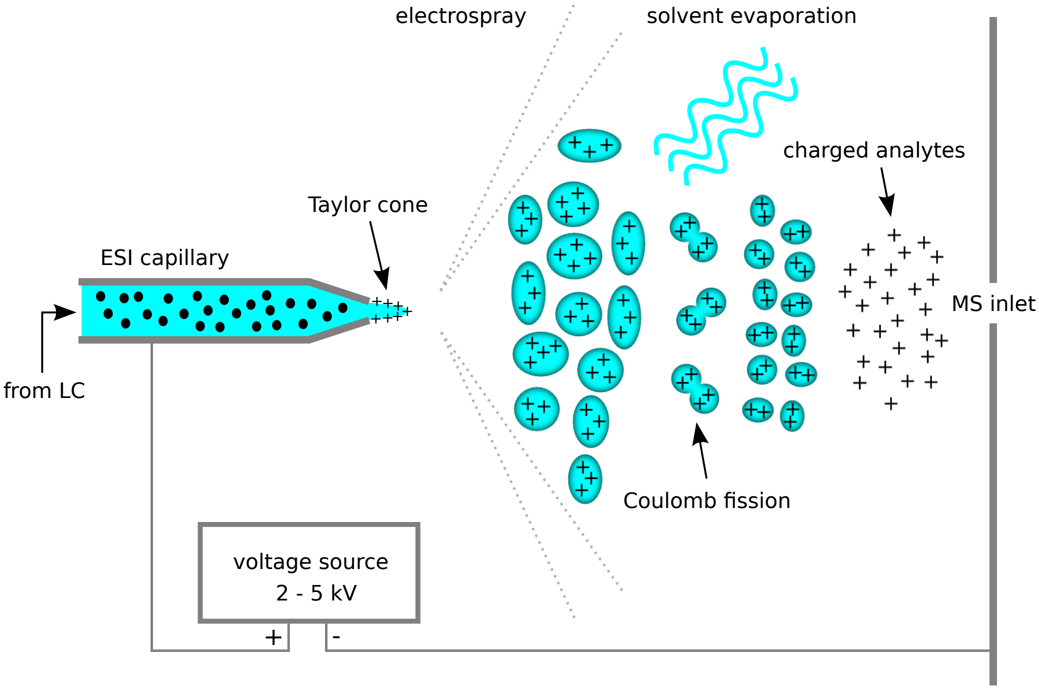


Figure 2.1: The electrospray ionization (ESI) method.

addition is the Orbitrap mass analyzer [34], a further development of the *Kingdon* [40] and *Knight ion traps* [41]. In a Kingdon trap, a central wire electrode is surrounded by a coaxial cylinder electrode. With voltage applied between the isolated electrodes, a radial electrostatic potential is built that can capture ions in a stable orbit if they enter the trap perpendicularly to the central axis and with the right velocity. The Knight trap introduced changes to the shape of the outer electrode such that the radial potential was extended by an axial quadrupole term, that is,

$$\Phi(r,z) = A \cdot (z^2 - \frac{r^2}{2} + B \cdot \ln r) \tag{2.1}$$

with r and z representing the cylindrical coordinates and A and B being constants reflecting the Kingdon trap geometry and voltage setup [34]. Neither the Kingdon nor the Knight trap were considered for potential applications in MS. Makarov further developed on these concepts to build a novel mass analyzer, the Orbitrap, with its applicability in MS in mind. Similar to the Knight trap, this mass analyzer builds an electrostatic potential that combines a radial logarithmic with an axial quadrupole component, that is,

$$\Phi(r,z) = \frac{k}{2}\left(z^2 - \frac{r^2}{2}\right) + \frac{k}{2} \cdot R_m^2 \cdot \ln\left[\frac{r}{R_m}\right] + C \tag{2.2}$$

with r and z being the cylindrical coordinates, k the field curvature, R_m the characteristic radius, and C a constant. Ions that are directed into the Orbitrap (Figure 2.2) adopt an orbiting trajectory around the inner electrode in r direction while oscillating in the z direction. The axial ion motion along z can be described independently from the radial motion as given by

$$z(t) = z_0 \cdot \cos(\omega t) + \sqrt{\frac{(2E_z)}{k}} \cdot \sin(\omega t) \tag{2.3}$$

where

$$E_z = \frac{m}{2} \cdot z_0^2 \tag{2.4}$$

corresponds to the potential of a simple harmonic oscillator and

$$\omega = \sqrt{\frac{q}{m} \cdot k} \tag{2.5}$$

to the frequency of oscillation along the z axis [42]. Determining this frequency ω facilitates to compute an ion's mass m and its charge q. This can be achieved by amplifying the induced image current on the split outer electrodes (Figure 2.2).

Q Exactive Orbitrap platform

Since the introduction of the first mass spectrometer based on the Orbitrap mass analyzer in 2005, the linear trap quadrupole (LTQ) Orbitrap, Thermo Fisher Scientific marketed several improved Orbitrap instruments. In general, these mass spectrometers were *hybrid* since they combined different types of mass analyzers in one device. For example, while the LTQ Orbitraps were prepended by a linear ion trap, the Q Exactive family employs a quadrupole mass analyzer as a mass filter [43]. Furthermore, the Orbitrap platforms differ in their ion optics responsible for the transfer of ions from the ionization source to the mass analyzer. We chose the Q Exactive benchtop platform as a new generation device to sketch the functioning of Orbitrap-based mass spectrometers (Figure 2.2).

Ions produced in the ion source are pushed through a *heated ion transfer capillary* to further support complete desolvation. The *radio frequency (RF) lens*, an array of thin plates with an RF voltage applied, focuses the ion beam and guides it through a metal aperture into the *injection flatapole*. This and the following *bent flatapole* are RF-only multipoles serving mainly for ion transportation. Furthermore, the bent flatapole's curvature facilitates a compact construction of the Q Exactive, but also filters out neutral molecules that may otherwise cause measurement noise. The ion beam reaches the quadrupole mass analyzer that allows to selectively analyze mass ranges (e.g., precursor selection for MS/MS) instead of being restricted to all-ion analysis [43]. After that, the filtered ion beam is guided by the following octopole into the *C-trap*. The C-trap, a curved linear trap, stores ions for specific time frames before releasing them into the Orbitrap mass analyzer. RF voltages on the trap's rods confine the ions

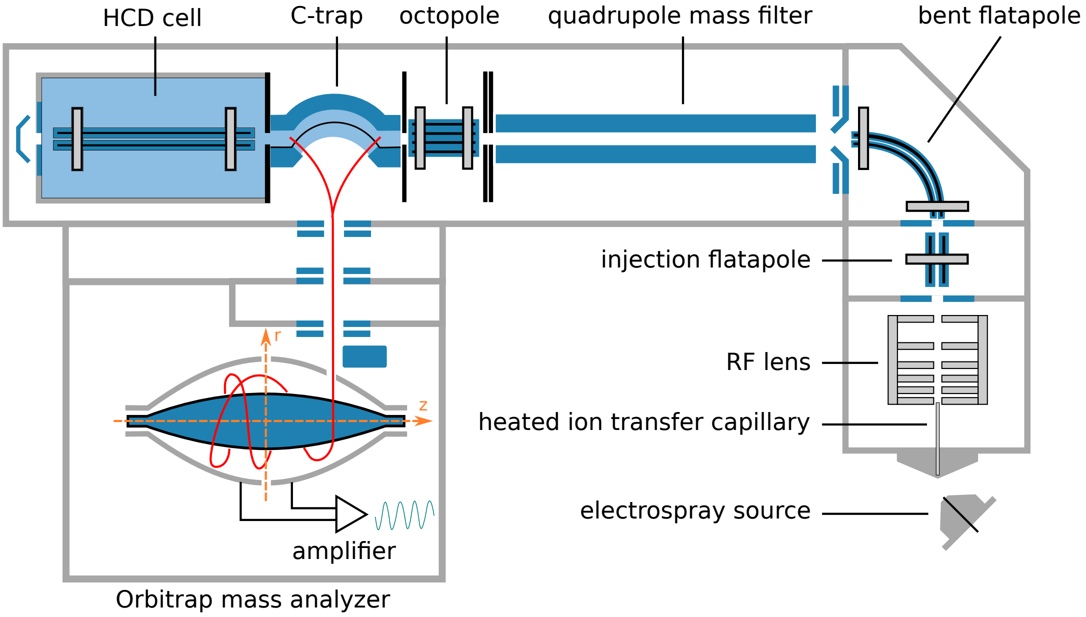


Figure 2.2: Schematics of Thermo Fisher Scientific's Q Exactive Orbitrap platform. This figure was drawn based on the schematics found in the operating manual for the Q Exactive [45].

to a back-and-forth movement along the longitudinal axis. Furthermore, the C-trap is filled with nitrogen to slow down ions with high kinetic energy incoming from the upstream octopole. Ions are delivered to the Orbitrap mass analyzer as *packets* by *pulsing* them in regular intervals. For this, the RF voltage along the axis is switched off and an extraction voltage is applied to transmit the ions perpendicularly to the trap axis. Intermittent lenses provide that the ions are focused and deflected to minimize gas carryover into the mass analyzer.

MS/MS Fragmentation

In case of an MS/MS experiment, the precursor ion of interest is first selected by the quadrupole mass filter, and then guided over the C-trap into the higher-energy collisional dissociation (HCD) cell [44]. This cell consists of a multipole embedded in a high-pressure gas chamber (e.g., helium or nitrogen). Upon entry, the precursor ions are accelerated along the multipole axis and fragmented through collisions with the inert gas molecules. The resulting fragment ions are redirected into the C-trap, from where they are pulsed into the Orbitrap the same way as described before.

2.1.2 Targeted versus Untargeted Metabolomics

Metabolomics experiments can be classified as either *untargeted* or *targeted* approaches [46]. In untargeted metabolomics, a mass spectrometric snapshot of a whole metabolome is taken to answer the biological question in interest. Since there is no prior knowledge about the compounds present in the sample, statistical methods are employed on all unknown metabolite signals to either classify the sample (e.g., healthy versus disease state) or to extract metabolic patterns characteristic for a specific physiological state (*metabolic fingerprinting* [46]). The identity of potentially interesting metabolites remains unknown until further experiments for their structural elucidation are conducted or computational methods are employed to annotate the unknowns with putative IDs (Chapter 4). For this reason, untargeted metabolomics can be regarded as the *hypothesis-generating* approach that is employed for *screening* or *discovery* purposes. In later chapters of this work, we present examples for such experiments (Chapters 5 and 6).

Contrary to the untargeted case, targeted metabolomics is *hypothesis-driven*. In this case, the analysis is directed towards a well-characterized subset of metabolites (e.g., key players in central metabolism) while other signals from the measurement are ignored. The metabolites' identity can be determined immediately with the help of authentic standards (e.g., with industrially manufactured standard kits, see Section 2.2.2). By focusing on specific metabolic pathways, direct interpretation of the data in the context of physiology is possible and thus particularly suitable to validate given hypotheses [46]. Although both the targeted and untargeted strategy begin with different premises, they can be combined to complement each other. For example, preliminary working hypotheses could be generated by an initial untargeted screening and then validated by more specific targeted experiments.

2.2 Computational Metabolomics

The technological advancement in MS facilitated the simultaneous measurement of hundreds or thousands of metabolites in a single experiment. The overwhelming complexity of such datasets increased the demand of efficient quantification algorithms and workflows to facilitate their processing in an automated and high-throughput manner. This bottleneck was addressed in recent years by active research and development of several computational strategies for quantification. However, the ID of these metabolites has turned out to be the next time-consuming step in untargeted metabolomics workflows.

2.2.1 Quantification

Quantification aims at the robust and reproducible extraction of the measured analytes' quantities from an MS experiment. It is usually the first step in a metabolomics workflow and thus its accuracy has a high impact on the quality of subsequent analyses. Analytical methods employed for quantification are coarsely categorized by *labeled* and *label-free* approaches. In general, labeling approaches introduce reference compounds into the sample that simplify the detection and quantification of target analytes. The labeled variants differ from their targets by a characteristic mass distance, but otherwise, behave chemically identical (particularly, show the same rt). The labeling process can be carried out in many ways, e.g., by adding stable isotopes as internal standards prior to the measurement [47], feeding the organism under study with stable isotope labeled nutrients [48], or targeting analytes with specific functional groups [49]. Based on the measured intensity ratio between the labeled variant and its target analyte, a relative or absolute quantification can be achieved. Although these techniques are appealing due to improved sensitivity in quantification, they also suffer from several drawbacks. Sample preparation is complex, time-consuming, and very expensive with regard to sample material and labeling compounds. In many metabolomics studies, sample material is usually very scarce (e.g., tissue biopsies) and may not suffice for the high demands of labeling approaches. Moreover, high purchase prices along with high consumption of labeling compounds (e.g., due to low efficiency of the chemical labeling process) raise the overall costs of metabolomics studies. For these reasons, labeling approaches might be feasible solely for small-scale metabolomics experiments.

In recent years, label-free quantification methods have been studied extensively as an alternative to labeling approaches. Simpler sample preparation protocols, reduced time and cost, better scalability, and easier to achieve results have made this alternative very attractive for large-scale proteomics as well as metabolomics experiments [50, 51]. A basic approach to quantifying analytes is first to detect their chromatographic peak profiles in a total ion chromatogram (TIC) and then to compute their peak areas, either manually or with computer assistance. In a classical differential analysis, these quantities can be matched between control and treatment groups in order to compute the analytes' *fold change* [50]. While this works well for samples with low complexity, the TIC becomes very difficult to analyze in case of complex biological samples (e.g., blood plasma with thousands of metabolites). This can be addressed by resolving the TIC in the m/z dimension in order to detect XICs; the peak area is then computed under the XIC profiles instead [9]. Here, the main difficulty for computational methods lies in the accurate detection and integration of chromatographic peaks since peak profiles are rarely Gaussian-shaped but rather exhibit heavy fronting and tailing [52]. Due to naturally occurring stable isotopes (e.g., ^{13}C and ^{15}N), MS resolves isotope patterns or *features* (series of coeluting ion chromatograms) rather than a single ion mass. We use the term *feature* to describe the total mass spectrometric signal that was caused by a specific analyte [11]. In order to reduce data redundancy and to increase the accuracy of quantification, feature detection algorithms are required to robustly detect the individual ion chromatograms (or *mass traces*) of an analyte's isotope pattern, reassemble them (deisotoping), and finally report a condensed representation of the feature (data reduction) [53].

Aside from feature detection, computational quantification methods usually comprise general signal processing (e.g., centroiding of profile mode data, noise and baseline reduction), correction of rt shifts [54], and linking features that coincide between LC-MS measurements to *consensus features* [55]. To this end, software packages have been developed that addressed most or all of these steps simultaneously in order to provide an all-in-one solution for computational quantification [9]. However, one severe drawback of such comprehensive tools is the missing flexibility when novel algorithms need to be integrated or existing components to be exchanged. This can be avoided if the individual processing tasks are implemented as self-contained modules, as in the case of the OpenMS framework [56]. The quantification strategy is then implemented as a *workflow* of such modules, offering both flexibility and robustness (Figure 2.3).

2.2.2 Identification

LC-MS-based metabolomics studies of complex samples (e.g., blood plasma) yield hundreds or thousands of signals that must be first annotated before any meaningful biological knowledge can be extracted. The efforts that have to be made for metabolite ID strongly depend on whether a targeted or untargeted approach is chosen. In a targeted approach, the analysis is restricted to a subset of metabolites that can be quickly identified with the help of reference standards. Manufactured assays facilitate the targeted quantification and ID of up to 190 metabolites that cover several compound classes [59]. Here, the investigation aims at testing specific hypotheses, for instance, if a pathophysiological condition is reflected by abnormal changes in one or more key metabolites from central metabolism [60].

In untargeted metabolomics experiments, however, no prior knowledge about the metabolites measured in a sample is available and there are usually only vague hypotheses to start with. Mass spectrometric signals with different intensities between two states (e.g., healthy versus diseased) can be detected with statistical methods as potential biomarker candidates. However, without any metabolite annotation, the biological interpretation of the experiment is delayed until further ID experiments have been conducted [61]. On the one hand, comprehensive annotation of hundreds or thousands of signals with experimental means is infeasible due to time and cost. On the other hand, although several computational methods have been developed to solve this problem in an automatic and efficient manner, metabolites identified merely by computational means are still regarded as unreliable. Thus, the untargeted approach is often employed as a *screening* or *discovery* tool to help with the design of more specific targeted metabolomics experiments.

The MSI proposed guidelines for reporting metabolite IDs in the literature [12]. Metabolite IDs are classified by their level of confidence: The more orthogonal data sources have been used for the ID process, the more reliable it will be considered. The classification levels proposed by the MSI are summarized in Table 2.1. Additional to the orthogonality criterion, the MSI guidelines require the confirmation of putative metabolite IDs through comparisons with authentic standards measured under identical conditions. For these comparisons, either authentic standard compounds must be spiked into the measurements or further MS/MS, MS^n, or even NMR experiments must be conducted. This validation process is often time-consuming and expensive (depending on the number of authentic standards to purchase) and thus most metabolomics studies are restricted to the ID of the most promising candidates.

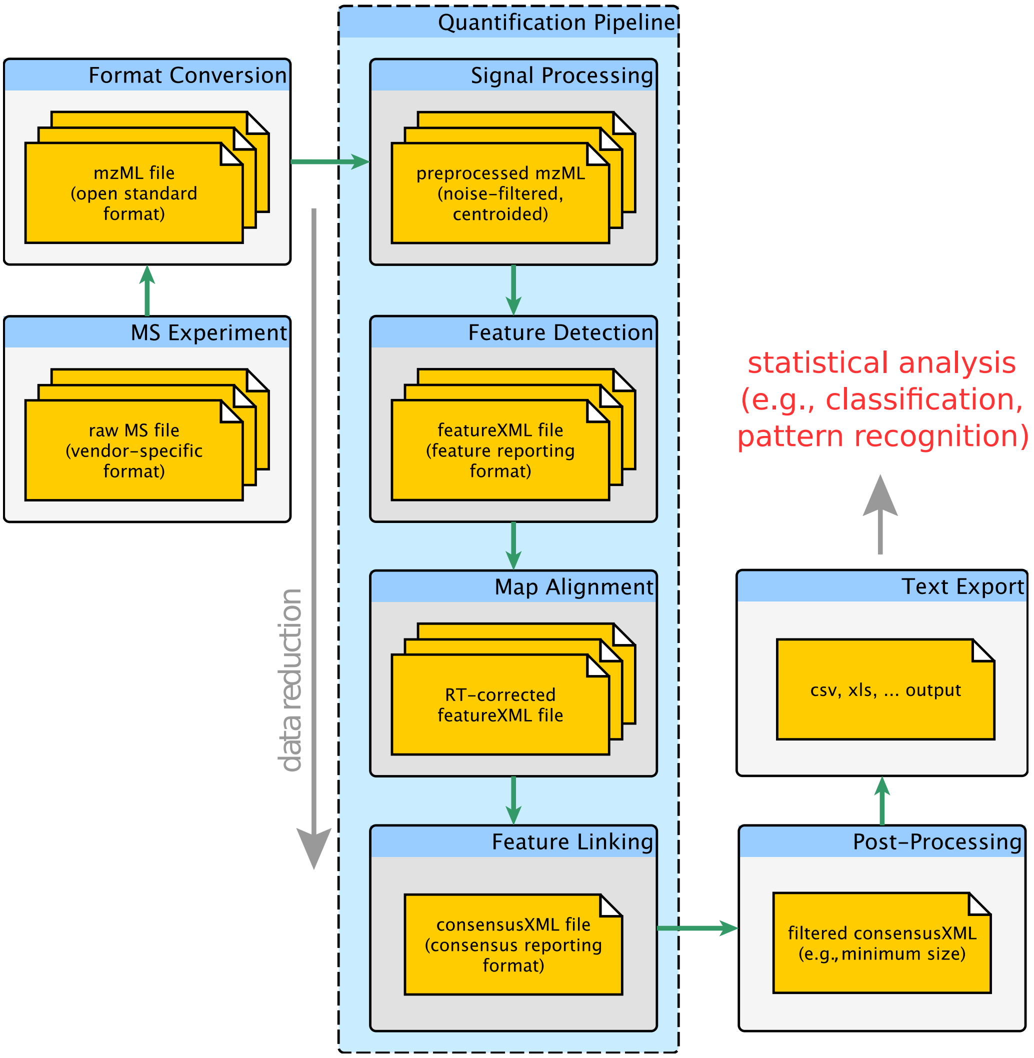

Quantification Pipeline
Format Conversion
mzML file (open standard format)
MS Experiment
raw MS file (vendor–specific format)
Signal Processing
preprocessed mzML (noise–filtered, centroided)
Feature Detection
featureXML file (feature reporting format)
Map Alignment
RT–corrected featureXML file
Feature Linking
consensusXML file (consensus reporting format)
data reduction
statistical analysis (e.g., classification, pattern recognition)
Text Export
csv, xls, ... output
Post–Processing
filtered consensusXML (e.g., minimum size)

Figure 2.3: General computational quantification workflow. Commonly, measurements are stored by the MS software in a vendor-specific binary format. These binary files must first be converted to the open data standard mzML [57] (e.g., by the Proteowizard software [58]) before the data processing and analysis with the OpenMS framework may commence. After data conversion, a typical quantification pipeline starts with signal processing (e.g., centroiding of profile data, noise and baseline reduction). Feature detection algorithms extract the essential analyte signals against the background of noise and store them in condensed feature lists (`featureXML` files [11]). With map alignment and feature linking, coinciding features from theses lists are merged into a consensus matrix (single `consensusXML` file). Both the `featureXML` and `consensusXML` files can be exported to tabular text that is readily readable by most statistical packages such as the R framework [8].

Level of confidence	Requirements
1 (confident)	at least two orthogonal criteria are in accordance to an authentic standard; spectral match to in-house DB of authentic standards (measured under identical conditions)
2 (putative ID)	physicochemical properties or spectral information match to external metabolite DBs; no authentic standards used
3 (putative class)	physicochemical properties or spectral information match to a specific compound class (e.g., phospholipids)
4 (unknown)	merely quantified metabolite features; interesting due to their differential expression

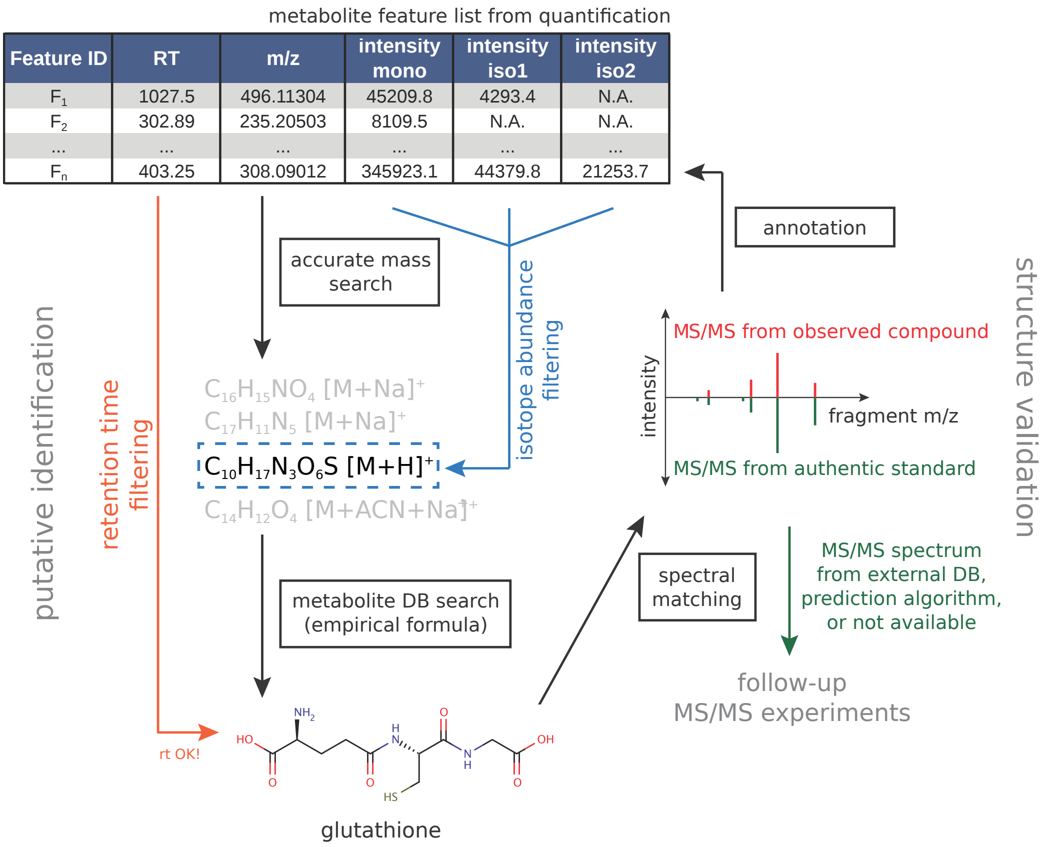
Table 2.1: The MSI guidelines proposed for metabolite ID. To confidently confirm a metabolite ID, the application of authentic standard compounds is essential.

According to the MSI guidelines, the bare minimum for confidence levels 1 and 2 is MS/MS information to match against spectral databases. ID workflows that are purely based on computational methods without MS/MS data are considered unreliable or unidentified. Consequently, the main purpose of such computational workflows is limited to filtering the number of false positive IDs and thus reducing time and costs of expensive follow-up experiments. A general outline of ID workflows for metabolite data is given by Figure 2.4 and parts of it are discussed in more detail in the next sections.

Accurate Mass Search

The first step in metabolite ID involves the calculation of empirical formulas from observed masses. An empirical formula represents the *elemental composition* of a molecule; in case of organic small molecules, the most common elements are C, H, O, N, P, and S. The problem of finding suitable empirical formulas for a given mass can be described as the NP-hard Money Changing Problem (MCP) [62]: Given a molecular mass M, a *weighted alphabet* containing k elements, and their associated atomic masses as weights $m_1, \ldots, m_k$, the problem is to find compositions (i.e., empirical formula strings) of non-negative integers $c_1, \ldots, c_k$ such that $M = \sum_i^k c_i \cdot m_i$. One classical approach to solve this problem is a dynamic programming (DP) algorithm which first precomputes a table data structure and then enumerates all empirical formulas matching the mass M [63]. However, the DP algorithm's running time and space requirements directly depend on the input size of M. More recently, an efficient algorithm for mass decomposition was proposed that is independent of input size M both in running time and computation space [64]. Nowadays, the automatic annotation of spectra with empirical formulas is considered as a standard feature in MS-related software packages.

The successful annotation of an unknown mass with its empirical formula highly depends on



metabolite feature list from quantification

Feature ID	RT	m/z	intensity mono	intensity iso1	intensity iso2
F_1	1027.5	496.11304	45209.8	4293.4	N.A.
F_2	302.89	235.20503	8109.5	N.A.	N.A.
...	...	...	...	...	...
F_n	403.25	308.09012	345923.1	44379.8	21253.7

accurate mass search
isotope abundance filtering
putative identification
retention time filtering
$C_{16}H_{15}NO_4$ $[M+Na]^+$
$C_{17}H_{11}N_5$ $[M+Na]^+$
$C_{10}H_{17}N_3O_6S$ $[M+H]^+$
$C_{14}H_{12}O_4$ $[M+ACN+Na]^+$
metabolite DB search (empirical formula)
rt OK!
NH2
HO
O
H
N
O
N
H
OH
O
HS
O
glutathione
annotation
structure validation
MS/MS from observed compound
intensity
fragment m/z
MS/MS from authentic standard
spectral matching
MS/MS spectrum from external DB, prediction algorithm, or not available
follow-up MS/MS experiments

Figure 2.4: Outline of a typical metabolite ID workflow. Based on the feature information from the quantification workflow (Figure 2.3), an accurate mass search determines all empirical formulas that match a specific feature mass within the given error margin. These formulas are then matched against a metabolite database to retrieve candidate structures. Orthogonal filter criteria such as rt and isotope abundance filtering are applied to further reduce the number of false positive candidates. In the final stage, the putatively identified structure is confirmed through spectral matching between the observed and the corresponding authentic standard's fragment spectrum.

its measurement accuracy. The advent of high-resolution mass spectrometers (e.g., FTICR or Orbitrap MS) promised to literally "read an analyte's empirical formula off its monoisotopic mass". However, it has been shown that even with a mass accuracy as high as 1 ppm the assignment of a unique empirical formula is only feasible for small molecules up to 200 Da [65]. The authors found that filtering out potential empirical formulas with non-matching relative isotopic abundances (RIAs) was a powerful orthogonal criterion that reduced the search space even more effectively than increasing the mass spectrometer's mass accuracy. In contrast to a significantly more expensive FTICR mass spectrometer with a below 1 ppm mass accuracy, a modern QTOF mass spectrometer (3 ppm mass accuracy and RIAs within 2 % root mean squared error (RMSE)) was adequate to assign unique empirical formulas even up to 300 Da when RIA filtering was applied. The likelihood of false positive annotations could be decreased even further by incorporating chemical knowledge into the filtering process. In their follow-up work, the same authors proposed a set of heuristic filtering rules commonly known as the *Seven Golden Rules* [66]. By removing empirical formulas that were unlikely in a chemical sense (e.g., showing unrealistic element ratios), these rules enabled the correct annotation with 98 % probability given that the empirical formula could be found in small molecule databases. With this example it became clear that sophisticated computational strategies exploiting orthogonal information from MS data offered a more powerful approach than relying on advances in mass accuracy alone.

Generally, it is assumed that the observed mass can be explained by a proton adduct $[\mathrm{M}+\mathrm{H}]^{+}$ in positive mode or a deprotonated ion $[\mathrm{M}-\mathrm{H}]^{-}$ in negative mode. Although these adduct ions are believed to be the most common ones, the formation of other adducts (e.g., with sodium or potassium ions in positive mode) is also observed frequently [67]. This must be taken into account when the alphabet for the mass decomposition is specified; otherwise, it could result in false annotations. Additionally, the observed mass might be the result of a double-charged ion (e.g., $[\mathrm{M}+2\,\mathrm{H}]^{2+}$). An alternative approach is to convert the adduct mass to a hypothetical neutral mass (e.g., in case of $[\mathrm{M}+\mathrm{Na}]^{+}$ by subtracting the mass of a sodium ion); the empirical formulas are then reconstructed based on this neutral mass instead. This neutral mass could also be directly queried against a metabolite database (for details, see Section 4.2.2). Contrary to a single adduct hypothesis, a set of several hypothetical adduct ions considered for a specific ion mass might also generate significantly more false positive IDs. The ionization behavior of small molecules in the ESI ion source is difficult to predict, but it is generally accepted that some adduct ions occur more often than others (e.g., $[\mathrm{M}+\mathrm{H}]^{+}$ versus $[\mathrm{M}+\mathrm{C_2H_3N}+\mathrm{H}]^{+}$). The computational annotation tool *MZedDB* incorporates ionization behavior rules to predict the most likely adduct ions formed and thus improves the annotation of unknown ion masses [68].

Apart from the formation of distinct adducts, a metabolite ion may be fragmented in the ion source and thus yield several distinct signals. Characteristic mass differences found in spectra reveal frequently occurring *neutral losses* (e.g., H_2O, CO, CO_2, or NH_3) [67]. These neutral losses are commonly caused by chemical transformations such as decarboxylation or deamination. This chemical relationship between two spectral signals can be exploited to further rule out false positive IDs [32, 69, 70, 71].

Structural Elucidation

Even if a unique empirical formula was found for a given mass, this information alone does not result in an immediate metabolite ID. Many potential metabolite structures may match the same empirical formula; for example, searching for C6H12O6 in the HMDB yields more than a dozen matching molecule structures [27]. On the other hand, C10H17N3O6S results in one single hit, namely glutathione (Figure 2.4). To resolve this ambiguity, it is required to incorporate structure-based information into the ID process. In chromatography-coupled MS, an analyte's retention behavior is dictated by its chemical structure. For this reason, its rt may serve as a powerful orthogonal criterion to filter out ambiguous IDs. While it is common practice to exploit the rt information in GC-MS experiments (usually by comparing *retention indices*) [72, 73, 74, 75], this is less well-established in LC-MS studies. Contrary to the high robustness and reproducibility of gas chromatography (GC), rt information obtained with LC techniques is much more variable due to changes in the manufacturing process of LC columns (e.g., differing stationary phases), varying elution gradients (e.g., differing solvent mixes), or external conditions (e.g., temperature). This is aggravated by the high chemical diversity of metabolites; for peptides, rt prediction models can be restricted to the small set of 20 amino acids with well-characterized physicochemical properties and thus are more accessible [76, 77]. Nonetheless, the recent development of rt models geared towards small molecules showed promising results. In case of hydrophilic interaction chromatography (HILIC), a multiple linear regression (MLR) model was presented that could predict rts on the basis of six physicochemical properties [78]. Applied as a filter criterion orthogonal to the accurate mass search, the number of false positive IDs could be reduced by up to 40 %. However, the model gave reliable predictions only for small molecules up to 400 Da. Molecules with molecular weights higher than 400 Da are expected to come from hydrophobic compound classes such as lipids where it is more common to use RPLC instead of the HILIC technique. Furthermore, a common essential requirement of LC-MS-based prediction approaches is that they are not directly interchangeable between labs and thus must be first measured and trained on a lab's individual setup. In case of CE-MS, an artificial neural network was presented to predict the migration times of cationic metabolites [79].

Although the number of false positive candidates can be reduced considerably by rt filtering, it does not guarantee a unique ID. Small molecules that are structurally distinct may still be identical in their elemental compositions and not significantly different in their rts. Such molecules can be found frequently in compound classes that contain unsaturated acyl moieties (e.g., phospholipids, mono-, di-, or triacylglycerids). For example, the unsaturated fatty acids *cis*-7- and *cis*-9-hexadecenoate differ in the position of their double bond but otherwise are identical in both their elemental compositions and molecular weights. Such ambiguities can be resolved through MS/MS. Here, the fragmentation of an unidentified ion is induced on purpose to gain further structural information by its fragmentation pattern. Depending on the type and degree of fragmentation, such patterns can be very characteristic for the underlying ion's structure. The isobaric compounds *cis*-7- and *cis*-9-hexadecenoate from our previous example can be clearly distinguished by their MS/MS spectra (Figure 2.5). A widely used fragmentation technique in MS/MS experiments is collision-induced dissociation (CID) [80]. With the advent of the modern Orbitrap mass spectrometers, higher-energy collisional dissociation (HCD) was introduced as a novel fragmentation technique (Section 2.1.1). Although very similar to CID in principle, it may produce significantly different fragmentation spectra.

The fragment masses and interjacent m/z intervals observed in an MS/MS spectrum are the result of specific gas phase fragmentation reactions. These can be interpreted by the chemist to draw conclusions about the underlying molecule structure. However, doing so is often tedious and time-consuming since typical LC-MS/MS experiments generate thousands of fragmentation spectra. This problem was addressed by efficient *spectral matching* algorithms that compare the spectrum of interest against a DB of known compounds. In classical approaches, the fragments observed in an MS/MS spectrum are compared against spectra from a fragment database, and the matching score is computed as the dot product of the matched fragment intensities [81]. Other algorithms employ scores that are more tailored to the compound classes of interest, e.g., the *HyperScore* [82] in the *X!Tandem* software [83] to score peptide fragmentation patterns or the *X-Rank* score geared towards small molecules [84]. In recent years, a huge effort has been made to build representative DBs with fragmentation data of metabolites. Common public fragment data repositories are MassBank [13] and METLIN [85]. Although these public resources assist greatly in the computational ID of metabolites (Section 4.2.3), they still suffer from problems such as incompleteness or data heterogeneity due to different LC-MS platforms and operating parameters (e.g., varying collision energies). Many research labs tend to build private in-house DBs with fragment spectra acquired from authentic standard compounds. However, the high cost of authentic standards and the acquisition of their spectra only allow for very small and specific subsets of metabolites. Furthermore, the willingness to make such valuable data public is rather low. To remedy the situation of incomplete metabolite fragment DBs, computational methods for the prediction of fragmentation patterns were proposed. For example, some methods precalculate fragmentation patterns for a large comprehensive chemical database and match these against the experimentally observed spectra [86, 87].

2.3 Statistical Methods in Metabolomics

Over the last years, several statistical methods were proposed for analyzing metabolomics datasets [88]. In general, the main aim of such methods is to detect distinctive metabolite patterns or biomarkers between sample groups (e.g., healthy versus diseased state) that help to understand the pathophysiological mechanisms and to develop novel marker-based diagnostics.

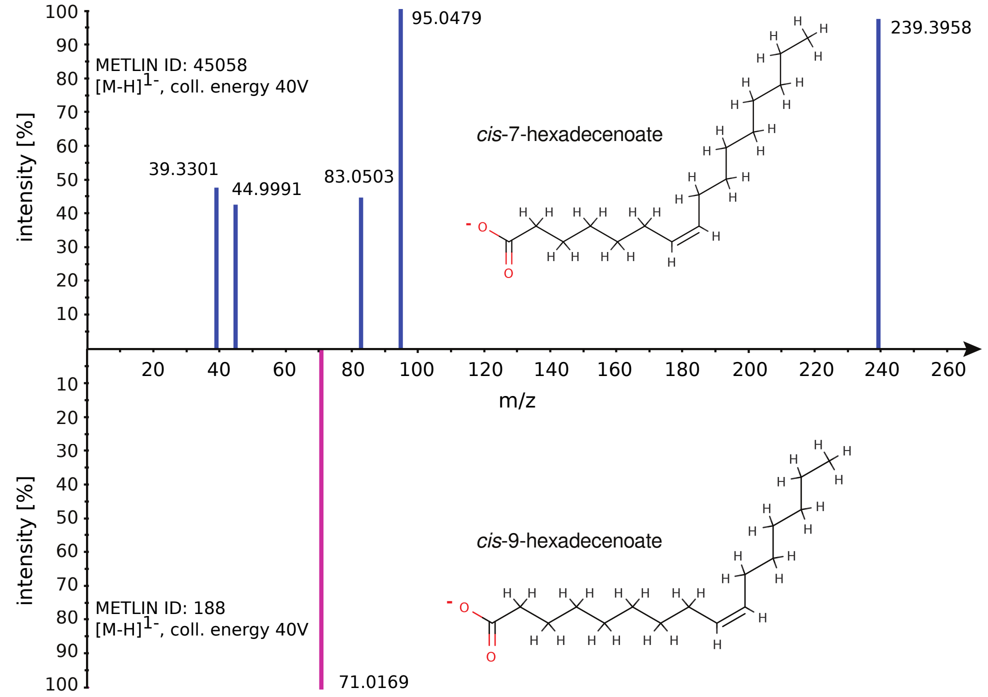


Figure 2.5: Comparison of *cis*-7- and *cis*-9-hexadecenoate with respect to MS/MS fragmentation. Both compounds (deprotonated) have the same elemental composition $C_{16}H_{29}O_2$ and with that the same molecular weight of 253.40092 Da. While it is not possible to distinguish these isomers by the means of an accurate mass search with orthogonal filter criteria, they exhibit characteristic differences in their CID fragmentation behavior.

The right choice of these methods is particularly important for the analysis of complex untargeted metabolomics datasets where we commonly have to cope with an enormous number of unidentified metabolite features. Differences in metabolite levels are often subtle (i.e., small fold changes between groups) and obfuscated due to strong biological variation within the groups (Chapter 6). This raises the need for proper pretreatment of the data (e.g., handling missing values, inter-sample normalization, data reduction) together with statistical techniques that are adequately robust and sensitive to detect subtle metabolite perturbations.

In a univariate setting, classical methods for biomarker detection comprise the t-test, Wilcoxon rank sum test, and ANOVA. These tests are applied to each metabolite feature independently and candidates are selected if they fall below a predefined significance level (e.g., $\alpha = 0.05$). For the most common metabolomics studies, one has to deal with the problem of *multiple hypothesis testing*. Metabolomics datasets (in particular, from untargeted analyses) comprise several dozens to hundreds of observations (samples) and hundreds to thousands of attributes (metabolite features) such that differences between groups might occur just by chance and result in a differentiation based on a few falsely accepted tests. To account for these errors in multiple testing scenarios, methods such as the *Bonferroni correction* or the concept of the false discovery rate (FDR) [16] were proposed.

In most cases, metabolic perturbations occur not isolated but as patterns of concertedly regulated metabolites. Since univariate methods do not take the dependencies between metabolite features into account, multivariate methods such as principal component analysis (PCA) and PLS-DA have been established as a quasi-standard for metabolomics biomarker studies. PCA is considered as a first-level exploratory tool to detect hidden structures in the data (e.g., if sample groups separate into distinct clusters). This method represents the data in a new orthogonal basis which is then rotated such that most of the data's variance is explained by a few principal components. For example, it was successfully employed to distinguish between urine profiles of controls and patients with renal cell carcinoma [89]. In contrast to PCA, PLS-DA is a supervised method that depends on a class information vector (e.g., control versus disease) in addition to the descriptor matrix. PLS-DA aims at transforming both matrices into a new space in order to reveal hidden structures in the descriptor matrix correlating well with the response. Due to the danger of overfitting, it is recommended to assess the predictive power of such model via cross validation techniques. In this context, it has been shown that the choice of the validation diagnostic dictates whether the final model favors low complexity or emphasizes metabolites with very small inter-group differences [17]. A PLS-DA model found its application in predicting kidney cancer on the basis of 30 metabolite markers [90]. In another study, the technique was employed for the discovery of biomarkers that were characteristic for a high-fat diet [15]. A further development of the partial least squares regression (PLS) concept is the orthogonal partial least squares regression (OPLS) model where the notion of noise is introduced [91]. The variation in the data is explained by two components: the variance correlated with the response and the orthogonal signal which is considered as noise. In recent years, OPLS-DA models have become popular in the metabolomics community due to a clear visualization and easy interpretability of their results. For instance, an OPLS-DA was employed to classify esophageal carcinoma with urine metabolite profiles [92]. However, some criticism has been raised about the practical use of OPLS-DA versus the classical PLS-DA [93].

Several novel statistical methods were developed to specifically meet the needs of omics studies. Although they were initially not designed for the analysis of metabolomics experiments, their concepts can easily be applied on such data as well. One of the earliest approaches to determine genes differentially expressed between sample groups was the simple computation of fold changes (FCs) and the application of arbitrary FC thresholds [94]. However, without any information about their statistical significance, one could not prove whether these reflected real differences or were merely observed by chance. The significance analysis of microarrays (SAM) was developed to detect statistically significant changes in gene expression measured with microarray technology and was made available as an R package [95]. Biomarker detection is achieved by correlating a test statistic (usually, the t-test) determined for each observation with the response and then computing their significance through repeated permutations of the data. As an example for its application in metabolomics research, the SAM was employed to detect key players in the metabolism of hepatitis C-infected tree shrews [96]. Another widely used R package implements the linear models for microarray data (limma) method [97]. In this approach, linear models are fitted to the expression profiles of each gene in order to explain the variability in the data. The method was also adapted for metabolomics studies, e.g., for the investigation of the metabolism in cystic fibrosis [98].

In the following sections, we discuss general strategies for the data pretreatment prior to the statistical analysis, i.e., intensity normalization and the filtering of irrelevant observations. We then detail the non-parametric statistical procedure of *rank products*, a robust tool to detect significant fold changes between the sample groups of interest [18].

2.3.1 Data Pretreatment Prior to the Statistical Analysis

Before any meaningful comparison between the samples' metabolite features can be conducted, the normalization of feature intensities is a crucial first step to reduce the technical inter-sample variation while preserving the biological differences. A popular method developed for microarray analysis is the *quantile normalization* [99]. It transforms each sample's individual feature intensities such that all the n samples share the same intensity distribution. In an imaginary quantile plot with n dimensions, the authors described this as a projection f_k of the p intensity quantiles $q_k = (q_{k1}, \dots, q_{kn})$ $(k = 1, \dots, p)$ onto the unit diagonal $d = \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$. The projection f_k for an individual quantile q_k has the form

$$f_k = \left(\frac{1}{n} \sum_{j=1}^{n} q_{kj}, \dots, \frac{1}{n} \sum_{j=1}^{n} q_{kj} \right) \tag{2.6}$$

i.e., the corresponding values of q_k in each of the n sample datasets are replaced by the mean quantile value. As an algorithm to compute the normalized intensity matrix X' from the matrix X of $p \times n$ dimensions, the authors proposed to sort the quantile values of each column, apply projection f_k to each row, and finally restore the original order of each column as it was before the sorting step. The algorithm is implemented and offered as an R package [100] and integrates well with downstream statistical analysis packages. Compared to conventional methods (e.g., total sum normalization), quantile normalization showed the strongest reduction of systematic errors while improving the data's interpretability in PCA analyses [101].

Consensus data matrices, as built by the feature linking step of our quantification workflow (Section 2.2.1), often show a high proportion of missing values. Such missing values may occur either due to failures of the quantification algorithm to detect features in the individual sample measurements or if the chromatographic peak intensity is below the level of detection (LOD). Consensus features with only few real values may originate from spuriously detected features (e.g., from noise or contaminants) and could hamper the subsequent statistical analysis if they are numerous. For this reason, special care should be taken when dealing with such irreproducible and (most likely) irrelevant metabolite features. In this regard, some filtering strategies were proposed for metabolomics data that aim at removing non-informative features and reducing the total number of missing values [15].

The first of the proposed filtering rules, the "80 % rule", drops all features that do not contain real values in at least 80 % of all samples. If the study involves more than one sample group, this rule is applied to each group, respectively. A feature is then dropped if none of the sample groups fulfills this requirement. In an experimental setup with two sample groups (e.g., ten control versus ten treatment samples), the group-wise application would avoid dropping a feature although the total number of real values happens to be low (e.g., eight real values exclusively in control group, $\frac{8}{20} = 0.4$). Such a situation may occur if a specific metabolite is detected reproducibly in the control group but is down-regulated upon treatment and thus falls below the LOD. The second proposed rule requires that quality control (QC) samples were measured throughout the experiment. Based on these samples, each feature showing a coefficient of variation (CV) higher than a predefined threshold (e.g., 30 %) is discarded. A suitable threshold value can be estimated through the technical variation observed for spiked standard compounds. This "CV rule" ensures that only metabolite features with reproducible intensities are passed to the statistical analysis. Since blank samples are measured routinely in the course of the experiment, it would be advantageous to leverage this information in another rule. In an ideal case, features showing up in the samples of interest should not be present in the blanks. However, due to carry-over effects, a trace amount of a specific metabolite could still be measured. Instead of testing the presence/absence of metabolite signals, a more robust rule should compare the median intensity level of the samples with the blank level. A feature is then discarded if the median sample level is not at least several folds higher than the median blank level. Such features often originate from contaminants.

For the metabolomics study of a T2DM-related polymorphism (Chapter 6), we implemented these filtering rules as vector functions in R [8] such that they can be applied directly on the rows (consensus features) of a normalized consensus matrix. This matrix must comprise all sample, blank, and QC measurements to facilitate an efficient computation (Figure 2.6). A consensus feature is accepted only if all three filter rules apply.

While the number of missing values in the dataset can be reduced tremendously by the presented filtering procedure, suitable data imputation techniques must be employed to deal with the remaining ones (e.g, regularized expectation maximization [15]). In general, most statistical software packages for biomarker discovery offer at least one simple data imputation technique.

2.3.2 Biomarker Detection with Rank Products

Rank products (RPs) were designed as a robust and intuitive statistical tool that makes rather weak assumptions about the data to analyze. In contrast to previous approaches, RPs do not presume anything about the data's error model, such as normality (t-test) or other symmetrical distributions of the error [18]. The method mimics the way a biologist would evaluate lists of genes ranked by their FCs: When a gene occupies the same or close ranks repeatedly over several replicate experiments, the probability of this observation just occurring by chance would

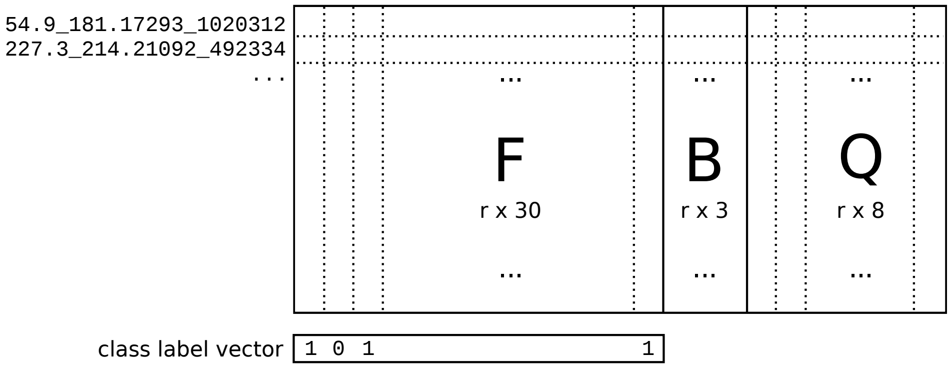


Figure 2.6: Consensus matrix layout prior to the statistical analysis. After coinciding features from the sample, blank, and QC measurements have been merged in the feature linking stage, the resulting consensus matrix is structured as illustrated here. Each row corresponds to one consensus feature and is uniquely labeled with the concatenated string of its centroid coordinates. The columns contain the feature intensities from the individual measurements. Submatrices *B* and *Q* facilitate the efficient application of row-based filter rules on the whole matrix (e.g., minimum signal-to-noise ratio between the individual and blank measurements). After the filtering operations, only submatrix *F* and the vector of the samples' class labels are submitted to the statistical analysis.

be very low. Given n observations and k replicates, this probability is $\frac{1}{n^k}$ if the ranking of the k gene lists were completely random. To express this likelihood, an RP is calculated for each gene g respectively by

$$\mathrm{RP}_g = \prod_{i=1}^{k} \frac{r_{g,i}}{n_i} \tag{2.7}$$

where $r_{g,i}$ is the rank of gene g and n_i the total number of genes in the list of replicate i. The computation of RP_g for up- and down-regulated genes differs such that each replicate's gene list must be sorted before by decreasing and increasing FCs, respectively. In case of single-channel arrays (e.g., Affymetrix chips) or quantification data from MS experiments, the FCs ratios between the two conditions to be compared must be computed in a pairwise fashion. Let m be the number of samples in both conditions C and T (e.g., control versus treatment). All pairwise comparisons yield the $m \times m$ FCs

$$\mathrm{FC}_{g,i,j} = \frac{T_i}{C_j} \quad i,j = 1,2,\ldots,m \tag{2.8}$$

where C_i and T_j corresponds to the expression/intensity level of control sample i and treatment sample j [102]. Then, a ranked list of FC ratios is generated for each of the $k = m \times m$ pairwise comparisons. The RP_g of a specific gene g is finally computed with the $\mathrm{FC}_{g,i,j}$ ratios from the k ranked lists.

In order to estimate the likelihood of observing a specific RP value and below by chance, a reference distribution is constructed by repeated permutations of the expression/intensity levels within the individual samples. In each permutation round, the list of RP'_g is recomputed on the shuffled data in order to refine the approximation of the reference distribution. From that, we can determine the expected frequency to encounter a specific RP value or below randomly with

$$\mathrm{E(RP)} \approx f(\mathrm{RP})/p \tag{2.9}$$

where p is the number of permutation rounds and $f(\mathrm{RP})$ the total count of occurrences ($\leq$ RP) accumulated over p rounds. The authors proposed the calculation of the percentage of false positives (PFP) that is conceptually similar to the FDR [16]. The PFP expresses how many false discoveries are to be expected if we consider all genes (sorted by increasing RP value) below this threshold as significant candidates. For each gene g, it can be calculated from its $\mathrm{E(RP}_g)$ by

$$q_g = \frac{\mathrm{E(RP}_g)}{\mathrm{rank_{RP}}(g)} \tag{2.10}$$

where $\mathrm{rank_{RP}}(g)$ corresponds to the rank of gene g within the RP-ordered list.

The functionality for an RP analysis was implemented as an R package [102] and thus is easy to integrate with the data preprocessing techniques as described before (Section 2.3.1). As input, the RP method in R accepts the normalized/filtered data matrix and a vector with sample group information. Furthermore, the number of permutations p can be configured for more accurate PFP estimations. Optionally, missing values may be imputed with the gene-wise mean of the existing expression/intensity values. The output is split into two tables containing the up- and down-regulated candidates alongside with additional information such as FCs, RPs, PFPs, and p-values.

CHAPTER 3

Automated Label-Free Quantification of Metabolites from LC-MS Data

Text and figures in this chapter were adapted with minor modifications from our work previously published in Molecular & Cellular Proteomics [53].

3.1 Introduction

Biological samples yield complex LC-MS datasets that usually contain up to hundreds of thousands of mass spectrometric signals [103]. Processing such complex datasets poses a challenge in experimental studies and, aside from metabolite identification, is considered as one of the major bottlenecks in computational metabolomics workflows [104]. For this reason, methods for the automated label-free quantification of metabolites have been studied extensively to meet today's experimental requirements. Since quantification algorithms are situated very early in computational workflows, downstream processing and analysis heavily depend on their outcome (Figure 2.3). Algorithms for metabolite quantification must be capable of extracting ion signals robustly from profile or centroided data against the background of measurement noise and finally yield a condensed table of compound-specific signal intensities [105].

The extraction of LC-MS signals that correspond to true metabolites is hampered by the non-specific nature of the ESI process, resulting in more than 90 % background signals in the spectra [106]. Due to this fact, the challenge for a feature finding algorithm is to identify all signals caused by true metabolites while avoiding the detection of false positives [107]. Detector noise, poor signal-to-noise ratios for low abundance metabolites, and the presence of numerous peaks from isotopes, contaminants, and in-source degradation products render this task difficult [105]. Moreover, metabolites can occur in multiple charge states and form different adducts in

ESI, giving rise to multiple features per analyte [67]. In recent years, several feature detection algorithms were developed that pursue different strategies to overcome these difficulties. These can be grouped into those that solely extract chromatographic peaks from the given raw data and others that additionally perform deisotoping, i.e., assembling subsets of chromatographic peaks into isotope patterns that most likely originate from the same compound.

A widely used tool for metabolomics data is the XCMS package implemented within the R statistical framework [9]. XCMS implements a binning strategy in the m/z dimension and with that avoids the problem of searching for peaks in m/z direction. Simultaneous feature detection and noise removal is then achieved in a bin-wise manner by a second derivative Gaussian matched filter. Other R packages are also available, such as apLCMS for LC-MS profiles with high mass accuracy [103]. Neither of them provides built-in functionality for isotope pattern assembly, which has to be added by other packages such as CAMERA [10, 108].

Although it has been argued that it is impossible to define optimal bins for all circumstances [109], binning in the m/z dimension is also performed by other tools such as MZmine [110] and MAVEN [105].

Software tools such as MEND, MapQuant, and MZmine identify chromatographic peaks by matching them with a particular model profile such as a Gaussian or an exponentially modified Gaussian [111, 110, 112, 113]. However, detection based on such predefined models may have low sensitivity since the shapes of elution profiles vary greatly from one compound to another and peaks that do not conform to the predefined shape will be discarded.

Aberg et. al. compare the two-dimensional (2D) representation of LC-MS data with the behavior of signals and noise on a radar screen. They implemented a Kalman filter for intra-sample tracking and alignment of mass spectra into pure ion chromatograms (PICs) [109].

In addition to the detection of chromatographic peaks, several tools such as MEND [111], MapQuant [112], and msInspect [114] pursue different deisotoping strategies. In msInspect [114], chromatographic peaks that coelute over time and show similar profile shapes are pooled together and are considered as isotopes of the same compound. Another strategy is fitting a three-dimensional (3D) model of a generic isotope pattern against the raw data and subtracting the fit from the signal [115]. In MapQuant [112], peaks are deisotoped by fitting isotope patterns to the observed 2D data.

In proteomics, a popular choice for many feature detection algorithms is to match the shape of peptide isotope patterns against an "averagine" model [116, 117]. However, for non-peptide compounds such as metabolites, the observed isotope patterns will not match, as they differ considerably in their chemical composition as assumed by the averagine model. Many available software solutions depend on numerous parameters that are difficult to interpret. These need to be optimized to obtain high quality results which often proves to be challenging as their influence on the tools' behavior is hard to predict for the user [90, 103].

In this work, we present the novel software tool FeatureFinderMetabo (FFM) for the label-free quantification of metabolites from LC-MS data. The algorithm detects chromatographic peaks in a robust and efficient manner without the need of binning the data in the m/z dimension. Since there are no assumptions regarding the chromatographic peak shape, low-intensity peaks are detected with high sensitivity. Furthermore, the algorithm comprises a new method for model-based deisotoping of metabolite LC-MS data that is based on SVM classification. This novel classifier is designed to capture isotopic abundances of organic molecules in general and thus is also applicable to the wide array of small organic molecules. We designed the algorithm to be configurable mainly by three intuitive parameters that reflect the characteristics of typical LC-MS data. The straightforward and intuitive configuration of our software tool allows researchers to achieve fast and high-quality results and to focus more on down-stream analysis instead. To assess the performance of our algorithm, we examined the influence of the adjustable parameters on a simulated benchmark dataset and the algorithm's quantification capability on a real-world dataset from human plasma samples. These samples were spiked with a set of standard compounds in a wide range of concentrations to reveal the relationship between the detected feature intensities and the actual quantities. We compared our algorithm's performance to that of XCMS in combination with the CAMERA package both on the simulated and the spiked human plasma data.

While designed with applications in metabolomics in mind, the algorithm is applicable to all small molecule data which means that it has applications beyond metabolomics (e.g., drug substance analytics, lipidomics, peptidomics, environmental analysis). Initial experiments also indicate that the algorithm deals well with data produced on separation technologies other than high performance liquid chromatography (HPLC), for example with capillary electrophoresis-mass spectrometry (CE-MS) data.

The presented algorithm is implemented as part of the OpenMS framework. The OpenMS proteomics pipeline (TOPP) and library were designed as a versatile and functional framework for developing MS data analysis tools, providing a rich functionality ranging from basic data structures to sophisticated algorithms for data analysis and visualization [56, 104, 118]. It is open source software and incorporates all steps needed for building powerful computational metabolomics workflows.

3.2 Algorithm

Our feature finder algorithm comprises two main stages, namely the mass trace detection and the feature assembly. In the first stage, signals in centroid LC-MS data that occur repeatedly over rt within a machine-dependent margin of mass error are gathered in a mass trace. Initially, a mass trace may contain signals from two different analytes (e.g., isobaric compounds overlapping in their elution profiles). To resolve this, the algorithm performs a filtering step to split these mass traces into individual chromatographic peaks. In the second stage, mass traces corresponding

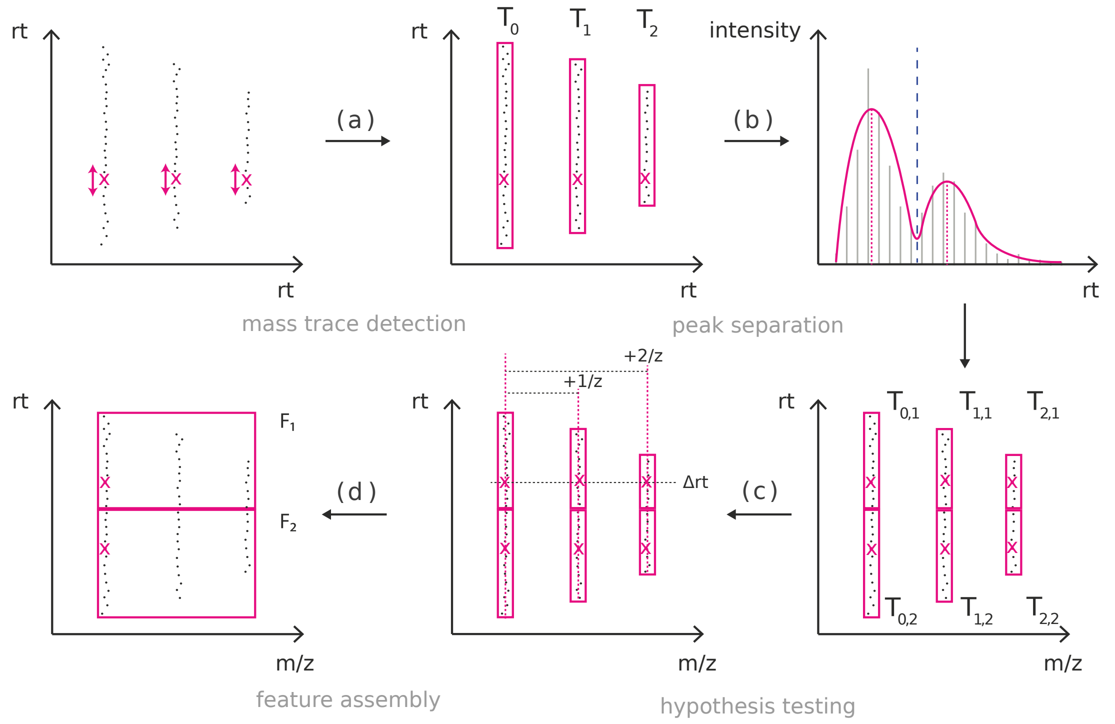


Figure 3.1: General procedure of feature finding. (a) Starting with the most intense peaks (magenta crosses), potential mass traces are extended with peaks compatible with respect to m/z back and forth in rt. (b) Each mass trace is smoothed to facilitate the determination of chromatographic maxima. Multimodal elution profiles are split into smaller mass traces with respect to the number of maxima. (c) Based on this set of mass traces, potential feature hypotheses are generated and scored by their compatibility with theoretical isotope patterns. (d) Finally, the best-scoring hypotheses are assembled into features.

to the same metabolite are assembled into *features*. Hypotheses sharing mass traces between distinct analytes are resolved by a scoring procedure. An overview of the algorithm is shown in Figure 3.1.

3.2.1 Mass Trace Detection

For mass trace detection, we assume that the LC-MS data has already been centroided (either by the instrument software or some other peak-picking algorithm). The data then consists of numerous spectrometric peaks $P = \{p_k\}$ where an individual peak p_k is defined by its rt t_k, mass-to-charge ratio m_k, and intensity i_k:

$$p_k = (t_k, m_k, i_k) \tag{3.1}$$

To prepare the input for the mass trace detection stage, we sort the set P by decreasing intensities and remove peaks that do not exceed a user-defined intensity threshold. Based on the resulting list P', the algorithm iterates first over the most intense peaks and considers each as a potential seeding point for the construction of mass traces. We define a mass trace T as a list of n mass spectrometric peaks $p_k \in P'$ that exhibit a similar m/z and occur in adjacent survey scans of an LC-MS run:

$$T = (p_1, p_2, \ldots, p_k, p_l, \ldots, p_n) \qquad t_k < t_l \quad \forall k < l \tag{3.2}$$

For each peak $p_k \in P'$, we initialize an empty candidate mass trace T with p_k as the seeding point. Starting from this seeding point, the algorithm attempts to extend T along the rt axis in both directions. This is accomplished by recruiting new peaks from P' that are close in m/z to all peaks gathered so far in T and thus originate most likely from the same ion mass. Depending on the accuracy of the underlying MS technology, mass spectrometric peaks exhibit scan-to-scan deviations from the true mass of the measured ion. These m/z errors follow a heteroscedastic noise model, i.e., low-intensity peaks are expected to have a less reliable mass than higher-intensity peaks [119]. We describe this error model by an intensity-weighted Gaussian distribution $\mathcal{N}$ with parameters μ (mean) and σ^2 (variance). The recruitment of suitable peaks to T depends on the accurate estimation of these parameters. To this end, we employ an online m/z density estimator [120] that refines the mean and variance estimates for each newly recruited peak p_{n+1} by the recursive expressions

$$\mu_{n+1} = \frac{w_n \cdot \mu_n + i_{n+1} \cdot m_{n+1}}{w_n + i_{n+1}}$$

$$\sigma^2_{n+1} = \frac{w_n \cdot \sigma^2_n + i_{n+1} \cdot (m_{n+1} - \mu_{n+1})^2}{w_n + i_{n+1}}$$

$$\text{with} \qquad \mu_n - 3 \cdot \sigma_n \leq m_{n+1} \leq \sigma_n + 3 \cdot \sigma_n \tag{3.3}$$

where $w_n = \sum_k^n i_k$ is the accumulated weights of the preceding trace peaks and $i_0 = 1$. Furthermore, we restrict the recruitment to those peaks that are most likely under the previously estimated m/z distribution $\mathcal{N}(\mu_n, \sigma_n^2)$. The recursion is initialized with a variance of σ_0^2 that is chosen proportionally to the expected mass error of the mass spectrometer. In most cases, this mass error turns out to be a rough estimate, since the online density estimator yields a more accurate estimate of σ^2 after only a few recursion steps and thus avoids erroneously recruiting potential outliers to the mass trace. Therefore, the algorithm is quite robust against incorrect choice of this user-defined mass accuracy parameter.

The extension of mass trace T aborts as soon as a specified number of scans is observed without finding an adequate peak. Such missing trace peaks occur more often in the fronting and tailing regions of a chromatographic peak as the mass trace starts to fade into the noise. We define a centroid m/z value for mass trace T by $\bar{m} = \mu_n$ where μ_n corresponds to the last mean estimate before aborting the trace extension. Trace peaks that were recruited by T are removed from the list P'. They can neither be used as a seeding point nor be part of other mass traces created in the subsequent iterations.

Analytes with mass-to-charge ratios sufficiently similar to be indistinguishable at the instrument resolution and sufficiently close rts can form overlapping elution peaks within a mass trace (Figure 3.1, upper right). Based on the elution profiles, these mass traces need to be split. The algorithm detects the chromatographic maxima of the respective elution profiles and minima between them. This is complicated by the fact that peak intensities along an elution profile tend to be noisy, in particular at low intensities. We use locally weighted scatterplot smoothing (LOWESS) with a polynomial of degree two, which offers the advantage that there is no need for a specific chromatographic peak shape to be defined in advance [121]. An important parameter that influences the degree of smoothing is the window size s, which describes the number of peaks the local polynomial fit is applied to. An optimal choice of s should roughly cover the extent of a chromatographic peak. We found that good estimates for the window size are the number of trace peaks that are covered within the full-width-at-half-maximum $\delta_{0.5}$ of a typical chromatographic peak. LOWESS smoothing is not very sensitive to changes in the window size. As long as the window size is reasonably close to $\delta_{0.5}$ of most chromatographic peaks — a parameter, which is generally well known for the chromatographic separation system — the smoothing will yield the desired result. Thus, we set the window size according to

$$s = \left\lfloor \frac{\delta_{0.5}}{t_{scan}} \right\rfloor \tag{3.4}$$

where t_{scan} designates the time between two consecutive MS scans.

The mass trace detection algorithm offers the option to automatically infer a good estimate of $\delta_{0.5}$ based on the data. To this end, we consider the distribution of peak widths that are gathered throughout the mass trace extraction phase. From this distribution, we can estimate an average peak width that is used for smoothing with good accuracy.

On the smoothed elution profiles, chromatographic peaks are detected. A mass trace peak is regarded as a local chromatographic maximum if it has a higher intensity than at least $\lfloor s/2 \rfloor$ trace peaks in each direction (increasing and decreasing in rt). This condition guarantees that, to be regarded as separable, the chromatographic maxima of two consecutive overlapping elution profiles are at least $\delta_{0.5}$ apart. If there are more than one chromatographic maxima within the same mass trace, the position with minimum intensity between each pair of maxima is determined. These minima are used as splitting points between two separable chromatographic peaks (Figure 3.2).

The splitting of mass trace T with c chromatographic maxima results in the c shorter mass traces $T_1, T_2, \ldots, T_c$. After the splitting step, we can define the rt for each mass trace T_j by the position of the chromatographic peak maximum. Additionally, we approximate the chromatographic peak area under the elution profile of a mass trace T_j by the Riemann sum

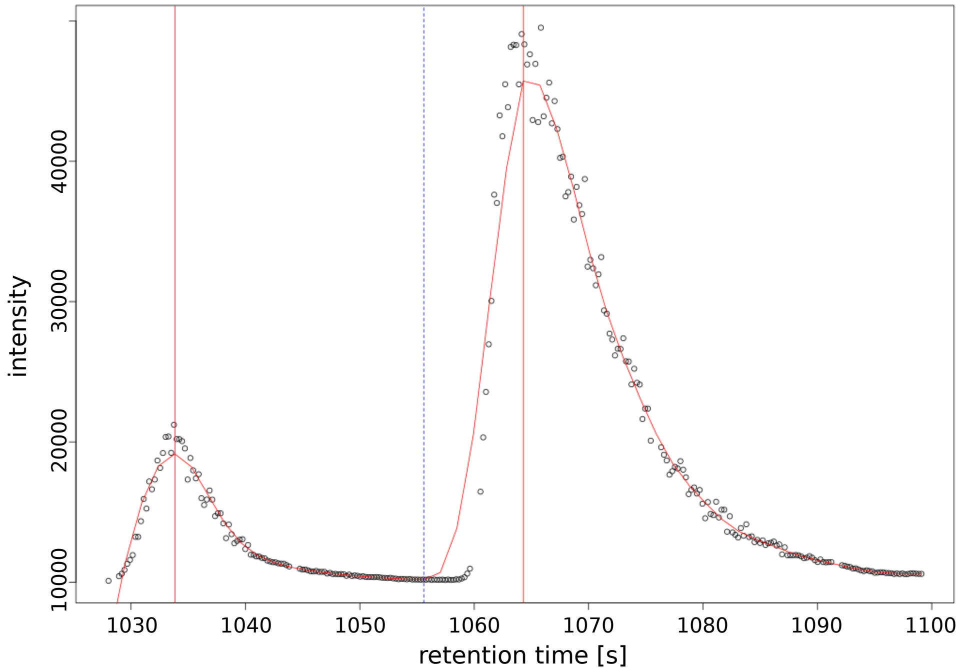


Figure 3.2: Separation of chromatographic peaks within the same mass trace. A smoothed multimodal mass trace (red line) is scanned for time points corresponding to local intensity maxima. Two vertical red lines indicate apex peaks that show higher intensity than at least $k/2$ neighboring peaks. The split into two separate mass traces is done at the time point with minimum intensity between these apeces (dashed blue line).

$$a_j = \sum_{l=0}^{n-1} i_l \cdot (t_{l+1} - t_l) \tag{3.5}$$

Finally, each individual mass trace T_j is added to a global list M that holds all detected mass traces for the following feature assembly.

3.2.2 Feature Assembly

The goal of the feature assembly stage is to find sets of mass traces that most likely originate from the same analyte. A set is regarded as a valid isotope pattern if its mass traces do co-elute, have correct m/z distances with respect to charge z, and exhibit correct isotopic abundance ratios (Figure 3.3).

The feature assembly algorithm reconstructs the isotopic patterns of multiple adduct ions of one analyte as individual features rather than aggregate them into one single feature. Following the modular approach of the OpenMS framework, such adduct features can be clustered subsequently or considered for improving the results of accurate mass queries against databases.

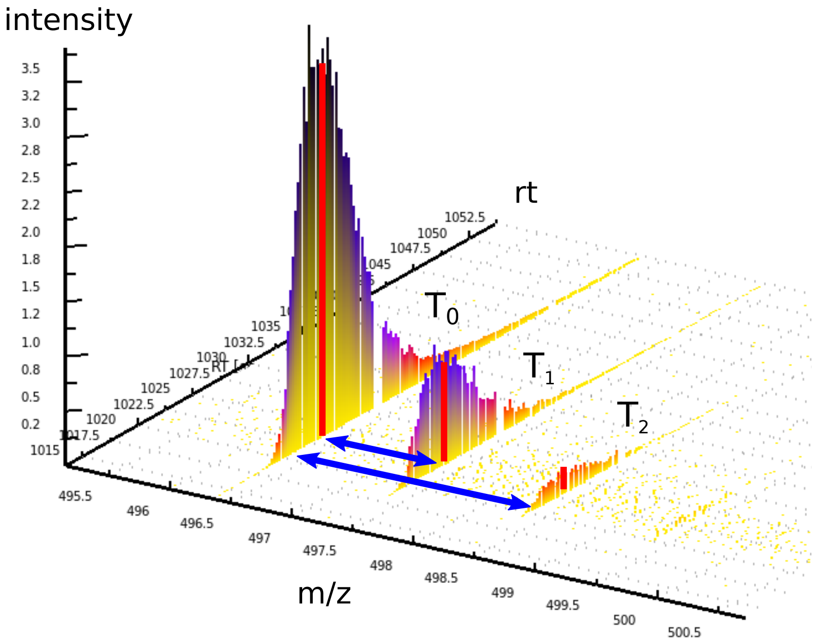


Figure 3.3: Characteristics of a metabolite isotope pattern measured in LC-MS. In a typical isotope pattern, mass traces belonging to the monoisotopic and higher isotopic ions can be observed. The relationships between these mass traces are characteristic for any naturally occurring isotope pattern: The m/z distances between each satellite (T_1, T_2) and the monoisotopic trace T_0 (blue arrows), the coelution of trace profiles (apeces indicated as red lines), and the intensity ratios between satellite traces and the monoisotopic trace.

We define a set of $k + 1$ mass traces as a feature hypothesis

$$H = (T_0, T_1, \ldots, T_k) \tag{3.6}$$

where the mass trace T_0 corresponds to the monoisotopic and $T_1, T_2, \ldots, T_k$ to higher isotopic mass traces of an analyte. To enumerate all potential feature hypotheses in a list C, the algorithm traverses all mass traces in M in order of ascending m/z. Each mass trace therein is iteratively assumed to be the monoisotopic trace T_0 of a candidate feature hypothesis H. After initializing H with T_0, the algorithm searches the list M for mass traces that are compatible with T_0 with respect to their rts, m/z distances, and intensity ratios. Instead of processing the complete list, we restrict the search for candidates to a subset of mass traces that lie in the vicinity of T_0 both in m/z and rt dimension. The criteria for m/z and rt compatibility are regarded as independent and are therefore modeled separately. Traces that fulfill these criteria are added to the feature hypothesis H.

In contrast to proteomics, there is no general model such as the averagine that could be employed as a filter to match m/z distances and abundance ratios of metabolite isotope patterns. Thus, we built a comprehensive set of isotope patterns characteristic for metabolites and studied novel models to validate our feature hypotheses H. While isotope patterns of typical metabolites could be obtained from metabolite databases, we chose a different approach in order to prevent potential biases of these databases, e.g., due to limited size. To generate a set of valid isotope patterns, we first employed the chemical formula generator program HiRes, which has been further developed by integrating heuristics to filter out unlikely sum formulas [66]. We generated all sum formulas based on the elements C, H, O, N, P, and S in a mass range between 1 and 1,000 Da. This yielded about 24 million distinct sum formulas. We randomly sampled a subset of 115,000 compositions spread evenly over the mass range of interest. We then computed the theoretical isotope patterns of these compositions with the program emass [122]. For each isotope pattern, we extracted the mass differences and abundance ratios between the monoisotopic and each of the higher isotope masses. The underlying distributions of these theoretical mass differences and abundance ratios provided the basis for novel isotope pattern models geared towards metabolites.

For each distribution of mass differences, we computed the mean and standard deviation. Based on the shifting elemental ratios towards higher masses and the distinct differences in mass between isotopes of different elements, we observed that the mass differences between isotope mass distributions were not constant but changed slightly and systematically, that is, they increased linearly from lower to higher isotopes. Instead of using fixed Gaussian models $\mathcal{N}(\mu, \sigma^2)$ for approximating each isotope mass spacing, we generalized the model generation by the linear equations

$$\begin{aligned} \mu_{\text{theo}}(j) &= 1.000857\,\text{u} \cdot j + 0.001091\,\text{u} \\ \sigma_{\text{theo}}(j) &= 0.0016633\,\text{u} \cdot j + 0.0004751\,\text{u} \end{aligned} \tag{3.7}$$

where $j = 1, 2, \ldots, 5$ corresponds to the higher isotope peaks considered. These models were obtained by linear regression based on the means and standard deviations of the theoretical m/z difference distributions (up to the fifth higher isotope), respectively (Figure 3.4). This generalization simplified the validation of feature hypotheses based on the mass differences δm between the monoisotopic mass trace T_0 and an arbitrary number of higher isotopic mass traces T_j $(j = 1, 2, \ldots, k)$. The m/z distance $\delta m(j) = |\bar{m}_0 - \bar{m}_j|$ corresponds to the difference of their m/z estimates computed during the mass trace detection stage.

To assess the similarity between the observed and theoretical m/z distances, we formulated the following error model: For each pair of traces T_0 and T_j with m/z variances σ_0^2 and σ_j^2 and the hypothetical charge z, the Gaussian models $\mathcal{N}_{\delta mz}(\mu(j), \sigma^2(j))$ with

$$\begin{aligned} \mu(j) &= \frac{\mu_{\text{theo}}(j)}{z} \\ \sigma^2(j) &= \frac{\sigma^2_{\text{theo}}(j)}{z} + \sigma_0^2 + \sigma_j^2 \end{aligned} \tag{3.8}$$

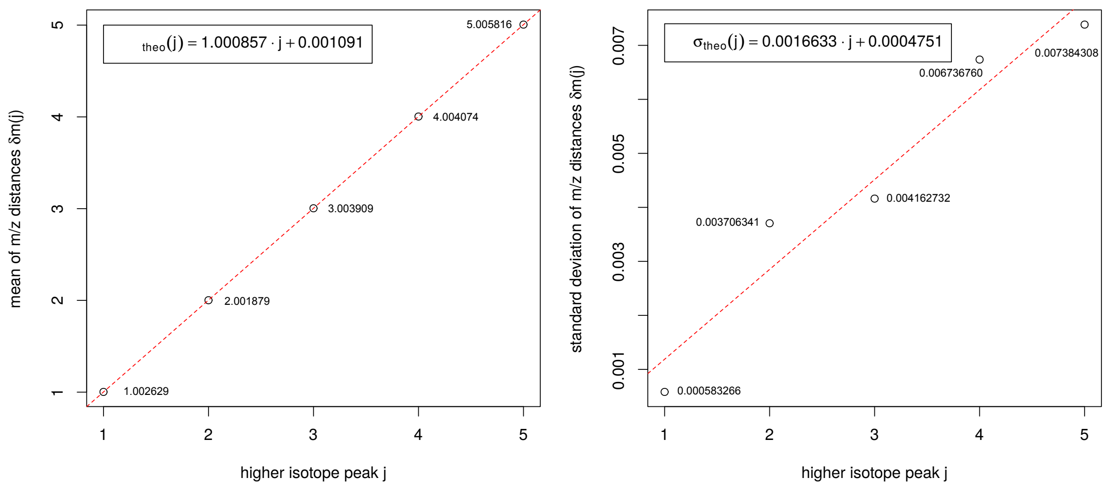


Figure 3.4: Linear regression models approximating m/z spacings within metabolite isotope patterns. The linear models (red dashed lines) were built on the basis of a large number of theoretical metabolite isotope patterns and their internal m/z spacings (difference in m/z between the monoisotopic and each of the higher isotope peaks). The linear fit was applied to the means and standard deviations of each of the respective m/z difference distributions.

are evaluated. This leads to the following scoring function for pairwise m/z distances:

$$S_{\delta m}(j) = \begin{cases} e^{\frac{-(\delta m(j)-\mu(j))^2}{2\sigma^2(j)}}, & \text{if } \mu(j) - 3 \cdot \sigma(j) \leq \delta m(j) \leq \mu(j) + 3 \cdot \sigma(j), \\ 0, & \text{else.} \end{cases} \tag{3.9}$$

In a nutshell, this scoring function assesses how likely it is that a set of mass traces was caused by the same metabolite based on a comparison to the precomputed mass difference distributions of potential metabolite compositions. It yields scores close to one for small mass differences and decreases with increasing deviation from $\mu(j)$. For values that lie outside the interval defined by three times the standard deviation $\sigma(j)$, the score becomes zero.

To ensure that mass traces compatible in m/z also exhibit similar elution profiles, we implemented a correlation similarity score as an orthogonal criterion. For each pair of mass traces T_0 and T_j, our algorithm first detects all peaks that are overlapping in both mass traces' full width at half maximum (FWHM) region. If this overlapping stretch comprises most of both FWHM regions (at least 70 %), we compute the similarity score or set the score to zero, otherwise. Given the mass traces' T_0 and T_j matched peak intensities (x_l, y_l), respectively, we compute their similarity with

$$S_{\mathrm{rt}}(j) = \max(\frac{\sum_l x_l \cdot y_l}{\sqrt{\sum_l x_l^2} \cdot \sqrt{\sum_l y_l^2}}, 0) \tag{3.10}$$

If the algorithm finds a suitable candidate mass trace T_j for $j = 1$, the scoring procedure is repeated to find the subsequent higher isotope traces for $j = 2, \ldots, 5$. We restricted the size of a feature hypothesis arbitrarily to a size of up to six isotopic traces (monoisotopic trace plus up to five satellite traces). As soon as the list of potential candidates is exhausted, we select the mass trace next in M as reference and construct all hypotheses supported by it. A candidate list C keeps track of all hypotheses that have been generated throughout this iterative process.

For each hypothesis $H \in C$, we compute a combined score S_{combined} with

$$S_{\text{combined}}(H) = \sum_{j=1}^{k \leq 5} \frac{1}{\sum^{N} a_j} a_j \cdot S_{\delta m}(j) \cdot S_{rt}(j) \tag{3.11}$$

In order to give preference to high-intensity signals over low-intensity ones, we weighed each score with the peak area a_j of mass trace T_j normalized by the total sum of the N detected mass traces in M.

We sort the candidate list C by decreasing score. Since C contains hypotheses built merely of mass traces compatible in rt and m/z dimensions, it is very likely that co-eluting or partly overlapping isotope patterns are merged into one single hypothesis instead of distinct ones. Furthermore, a trace that is actually a higher isotopic trace of another pattern might erroneously be chosen as the monoisotopic trace of a hypothesis. To filter out such invalid hypotheses, we first consider each hypothesis by decreasing score, giving precedence to high-scoring hypotheses. Subsequently, we apply an isotope abundance filter to reject unlikely metabolite hypotheses.

Based on the theoretical isotope abundances extracted before, we trained a support vector machine (SVM) to distinguish typical metabolite patterns from unlikely intensity ratios. For each of the roughly 115,000 isotope patterns, we extracted the first three isotope abundance ratios together with the monoisotopic mass. These constituted the four features of a training instance for the SVM classifier. Since the classifier must cope with isotope ratios from real-world measurements that are generally prone to errors, it might misclassify measured ratios when it is trained merely on theoretical isotope abundances. Consequently, we added Gaussian noise to each of the theoretical abundance ratios in order to model a RMSE of 5 %. In case of modern TOF mass spectrometers, an RMSE of 2 to 5 % is usually reported for relative isotopic abundances (RIAs) [66]. We validated our binary classifier on a separate test dataset based on empirical formulas from HMDB (around 7,800 metabolites with masses below 1,000 Da) and an equally sized set of randomly sampled negative test samples. The classifier achieved an excellent correlation between the actual and predicted classes (Matthews correlation coefficient (MCC) [123] of 0.974).

With the SVM classifier, we evaluate each feature hypothesis $H \in C$ and discard those that are classified as invalid from C. In the final step, our algorithm traverses C sorted by descending hypothesis scores S_{combined} and subsequently transfers feature hypotheses to a list of accepted features. In order to ensure that no mass trace occurs in more than one feature, all remaining hypotheses containing already accepted mass traces are removed from C. Due to the monotonicity of our scoring function, a feature hypothesis with $n \geq 1$ higher isotope traces will always yield a lower score than the same hypothesis extended by one or more higher isotope traces (both hypotheses validated by our SVM classifier). As a consequence, our algorithm will favor feature hypotheses that show the most complete isotope patterns and discard any of their subcomponents, maximizing the overall score and yielding the optimal solution.

The final list of features is stored in an extensible markup language (XML)-based file format (featureXML) that contains both key properties (rt, m/z, intensity, FWHM, and charge) and potential meta-information (e.g., unique identifier). The algorithm offers the option to report the feature intensity either as the chromatographic peak area of the monoisotopic trace alone (default setting) or the sum of all isotopic peak areas.

3.3 Datasets

Assessing quantification algorithms is a difficult task, severely hampered by the complexity of the data and the lack of an exact ground truth if experimental datasets are used (e.g., due to the presence of unknown/unexpected contaminants). At the same time, there is not a single performance measure for a quantification algorithm. It needs to be sensitive, accurate, fast, and robust. In order to assess the different aspects of our algorithms, we devised two complementary benchmarking strategies. For the first benchmarking setup, we conducted MS measurements with a concentration series of standard compounds that were spiked in human plasma. The resulting dataset allows us to assess the algorithm's quantification accuracy on a complex real-world sample. However, the exact number of true metabolite features in our experimental data is not known and thus the classical assessment by sensitivity and specificity is infeasible. Since there are only few freely available and fully annotated metabolomics datasets that could be used as a ground truth, we computationally generated a synthetic dataset based on a published list of 500 well-characterized plant metabolites. This ground truth provides the basis for our second benchmarking setup where we aim to quantify the algorithms' performance by terms of recall and precision rates.

3.3.1 Quantification Dataset

For the spike-in experiments with human plasma, we selected a set of seven standard compounds that provide both a reasonable distribution within the experimental gradient and a good mass coverage. The standard compounds together with their mass-to-charge ratios and rts under the given experimental conditions are shown in Table 3.1.

Standard	CAS-RN	m/z	rt [min]
Propionyl-L-carnitine-d3	1182037-75-7	221.15751	1.4
Nialamide	51-12-7	299.15025	5.7
Sulfadimethoxine-d6	73068-02-7	317.11851	8.4
Reserpine	50-55-5	609.28065	10.6
Terfenadine	50679-08-8	472.32099	12.4
Hexadecanoyl-L-carnitine-d3	1334532-26-1	403.36096	16.6
Octadecanoyl-L-carnitine-d3	N.A.	431.39227	18.6

Table 3.1: Standard compounds employed in the spike-in experiments.

Sample Preparation

All following sample preparation steps were performed by Sara Forcisi and Kilian Wörmann (Research Unit Analytic Biogeochemistry, Helmholtz-Zentrum München, Germany).

All aqueous solutions were prepared using LC-MS grade water from Chromosolv®, Fluka Analytical (Sigma-Aldrich, Munich, Germany). LC-MS grade acetonitrile and methanol were purchased from Chromosolv®, Fluka Analytical (Sigma-Aldrich, Munich, Germany). Formic acid UPLC-MS grade was procured from Biosolve (Valkenswaard, The Netherlands) and Sodium hydroxide (≥98 %) from Roth (Karlsruhe, Germany). Leucine enkephaline (≥95 % HPLC grade), Nialamide (95 %), Reserpine (≥99 % HPLC grade), Sulfadimethoxine-d6 (VETRANAL™), and Terfenadine were purchased from Sigma-Aldrich (Munich, Germany), [3,3,3-d3]-propionyl-L-carnitine hydrochloride, [16,16,16-d3]-hexadecanoyl-L-carnitine hydrochloride, and [18,18,18-d3]-octadecanoyl-L-carnitine hydrochloride were purchased from Dr. H. J. ten Brink Laboratory (Amsterdam, The Netherlands). A stock solution for each standard was prepared at the concentration of $1\,\mathrm{g\,L^{-1}}$. Each standard was mixed at the same concentration in 20 % acetonitrile solution. An experimental standard serial dilution at $10\,\mathrm{mg\,L^{-1}}$, $5\,\mathrm{mg\,L^{-1}}$, $2\,\mathrm{mg\,L^{-1}}$, $1\,\mathrm{mg\,L^{-1}}$, $0.5\,\mathrm{mg\,L^{-1}}$, $0.2\,\mathrm{mg\,L^{-1}}$, $0.1\,\mathrm{mg\,L^{-1}}$, $0.05\,\mathrm{mg\,L^{-1}}$, $0.02\,\mathrm{mg\,L^{-1}}$, and $0.01\,\mathrm{mg\,L^{-1}}$ was set. In addition, a non-spiked sample was included in the experiment.

Fresh EDTA blood was collected from healthy male donors. Blood plasma was prepared via centrifugation at 2,000 g at 4 °C for 7 min. The resulting plasma was pooled, mixed, and aliquoted for storage at −80 °C until analysis. Frozen EDTA plasma was thawed on ice and vortex-mixed for 30 s prior treatment. The protein precipitation extraction procedure was performed adding cold acetonitrile (1 mL) to a plasma volume (250 µL). After adding the organic solvent, the samples were vortex-mixed for 30 s at room temperature and centrifuged at 15,294 g for 10 min at 4 °C. The dried samples were reconstituted in 50 µL of the different experimental standard solutions.

UPLC-MS Analysis

The UPLC-MS analysis and data conversion steps were performed by Sara Forcisi and Kilian Wörmann (Research Unit Analytic Biogeochemistry, Helmholtz-Zentrum München, Germany).

Sample analysis was performed using a Waters Aquity UPLC system coupled to a Synapt HDMS oa-QTOF mass spectrometer (Waters, Mildford) equipped with an ESI operating in positive mode. The gradient chromatographic separation was performed with a C18 Vision HT-HL UPLC column (2 × 150 mm, 1.5 µm, Alltech Grom GmbH, Germany). Elution buffer A was water containing 0.1 % formic acid and elution buffer B was acetonitrile. The flow rate was set to 0.3 mL min^{-1}. The linear gradient method consisted in 5 % of B over 0–1.12 min, 5 to 100 % of B over 1.12–22.27 min and held at 100 % B until 29.49 min, returned to 5 % of B at 29.56 min. In order to equilibrate the column with the initial mobile phase, 5 % B was kept until 35 min. The column oven was set to 40 °C and the sample manager temperature to 4 °C. A solution of leucine enkephalin (556.2771 u, 400 pg μL^{-1}) in MeOH/H_2O:1/1 containing 0.1 % of formic acid was infused as lockmass compound at a flow rate of 5 µL min^{-1}. The spectra were acquired in centroid mode within an m/z range of 50–1,000. The detection parameters are listed in table A.1.

The samples were measured in randomized triplicates and alternated with blank samples consisting of 20 % acetonitrile. All data files were converted into mzML format [57] using the command line tool `msconvert.exe` of the ProteoWizard suite (version 2.1.2464) [58].

For each measurement, we extracted metabolite features with our algorithm (see Appendix A.1.1 for the configuration file) and stored the resulting feature map in the featureXML data format [11]. For the following processing steps, we used existing software solutions from the OpenMS framework (version 1.10). Features coinciding in at least two measurements were aggregated to consensus features and stored in the consensusXML format by using the `FeatureLinkerUnlabeledQT` software. This consensus file was then exported to a tabular text format with the `TextExporter` tool. All further analyses were conducted with the MATLAB (version R2012b) [124] and `R` (version 2.15.2) [8] software packages.

3.3.2 Simulated Plant Metabolites Dataset

In order to determine precision and recall of our methods as well as the dependence of the algorithm's performance on noise and instrument accuracy, we chose to employ simulated datasets. In case of the human plasma dataset, the computation of such metrics is infeasible since the ground truth is unknown. Thus, we created a dataset with known ground truth that is based on a metabolomics investigation of plant metabolites [125]. However, elution profiles and isotope patterns were created synthetically in order to control these parameters independently, which is not easily done in an arbitrary fashion experimentally. Even if it were varied experimentally, for example by repeated measurements modifying instrument resolution, we would not be able to remove inter-sample variance. The simulation also does not contain unknown contaminants so we can determine true recall and precision rates based on the known (simulated) ground truth.

For the generation of synthetic LC-MS datasets, we employed the software package `MSSimulator` (version 1.10) [126]. The simulator creates profile data based on the experimentally determined rt and the composition of a set of compounds. It can simulate different instrument resolutions, noise levels, and chromatographic performance. The datasets created here are based on experimentally identified metabolites and thus capture the complexity of a real-world sample. At the same time, the simulator provides complete control over signal-to-noise ratio, accuracy, and similar parameters that would be very difficult to vary independently of each other experimentally.

The dataset is based on a list of identified plant metabolites of a rather comprehensive study by Giavalisco et al. [125]. The authors reported the elemental composition, intensity, and rt of hundreds of metabolites in their work in the supplementary information of their study. This information was sufficient to simulate synthetic datasets with characteristics typical for a complex metabolomics study. `MSSimulator` performed a raw signal simulation of 500 metabolites within the m/z range of 100–1,000 and with a simulated gradient runtime of 25 min. The machine-dependent settings of `MSSimulator` were chosen close to our experimental MS setup: The simulated MS resolution R was set to 20,000 and scan time to 0.25 s. The output of `MSSimulator` comprised a centroided LC-MS measurement and a feature map summarizing all simulated metabolites with theoretical rts, m/z ratios, and intensities (Appendix A.1.1). This feature map facilitated the comparison between the ground truth (the simulator's input) and the features detected by our algorithm applied to the same simulated data.

To assess the influence of different sources of noise on the feature finding performance, we conducted three series of MS simulations each affected either by detector noise, m/z variation, or elution profile distortion. In the first series, we simulated increasing levels of detector noise that was controlled by the Gaussian standard deviation σ_{det}. The second series comprised two datasets simulated with moderate and strong mass errors (Gaussian standard deviation $\sigma_{m/z}$ set to 10 ppm and 40 ppm respectively). For the third series, we simulated elution profiles with ideal peak shape and heavily distorted shape. In `MSSimulator`, the degree of distortion is controlled by the number of iterations n_{dist}, i.e., the more iterations are performed the more the elution profiles deviate from their smooth Gaussian shape. The configuration for each simulation is summarized in Table A.2 and can be found under Appendix A.1.1.

For performance evaluation, we computed both the recall r and precision p based on the number of ground truth features and the number of features found or missed by our algorithm. These measures can be combined into the F-score [127] defined by

$$F\text{-score} = \frac{2 \cdot r \cdot p}{r + p} \tag{3.12}$$

For each of the simulated datasets, we employed our algorithm to detect a feature map and compute its F-score. The resulting scores were plotted with respect to either varying `MSSimulator` parameters (e.g., σ_{det}) or parameters of our feature finding algorithm.

Comparison to Related Methods

We repeated the performance assessments done with our algorithm with the XCMS and CAMERA packages (Bioconductor version 1.34 and 1.14, respectively) [9, 10], which are currently considered to be among the leading feature detection algorithms. For each experiment, we adjusted the corresponding parameter settings of the XCMS software to the best of our knowledge. Since the feature set found by XCMS is not deisotoped and thus would not allow for a fair comparison, we additionally employed the CAMERA package to detect higher isotope masses in the XCMS output.

To allow for a direct comparison to our method, we first converted the XCMS/CAMERA output to a compatible feature map. For each set of annotated isotope traces in the XCMS/CAMERA output, we kept the rt, m/z, and intensity values of its first isotope in a tabular format and discarded the higher ones. Since CAMERA clustered multiple adducts belonging to the same compound together, we resolved these clusters and stored each contained feature individually. Non-annotated features were regarded as singleton mass traces and were added to the same table. This extraction procedure resulted in condensed feature lists that were converted to the featureXML file format with the `FileConverter` software [11]. We employed the `FeatureLinkerUnlabeledQT` software to link features that were detected both by our algorithm and XCMS/CAMERA.

3.4 Results

In this section, we present the results obtained from the quantification dataset as introduced in Section 3.3.1. We assessed the performance of our algorithm from different perspectives. First, we examined the linearity of the response observed in spike-in experiments as a measure of quality of absolute and relative quantification. Then, we assessed the reproducibility of the feature finding process by determining the number of features detected over technical replicates. In order to dissect some effects of the data quality on the behavior of the algorithm in more detail, we show results obtained on a simulated dataset. With the general quality of the methods already demonstrated on real-world data, simulated data permit the independent variation of certain characteristics and also the determination of an independent ground truth. Finally, we repeated the benchmarking studies with the state-of-the-art software packages XCMS and CAMERA and compared the results to the performance of our algorithm.

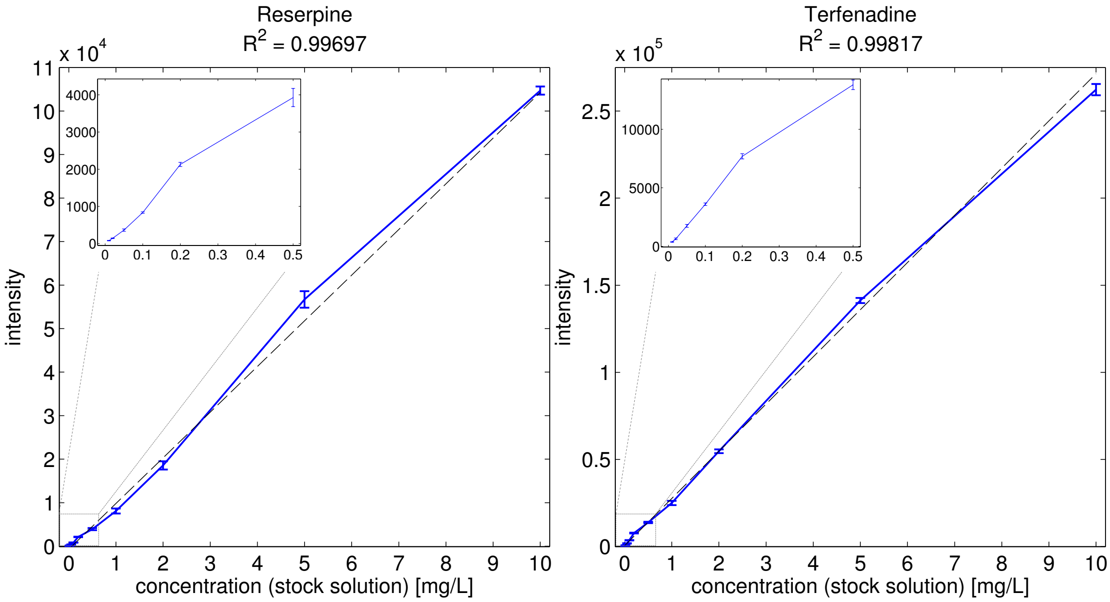


Figure 3.5: Relationship between Reserpine and Terfenadine concentrations and respective feature intensities. For each concentration, an error bar depicts the standard deviation of the triplicate intensities. The solid line connects the mean intensities obtained from each of the 11 triplicates. For each compound, a line was fit by linear regression to the 33 concentration-intensity data points with the goodness-of-fit R^2 (dashed line). The inlaid plots show a close-up of the low-concentration range between 0 and 0.5 ppm.

3.4.1 Quantification Linearity and Reproducibility

Based on the spike-in series in a complex plasma background, we assessed both the linearity (11 different concentrations spanning three orders of magnitude) and the reproducibility of the quantification (based on triplicate measurements). With the detection of the standard compound features for each of the human plasma spike-in experiments and their subsequent linking, we obtained intensities for each analyte and dataset. Table A.3 lists the corresponding mean feature intensities and standard deviations across the triplicate measurements performed for each concentration of the analytes. We show the relationship between the analytical concentrations and their corresponding feature intensities for two selected standard compounds (Reserpine and Terfenadine) in Figure 3.5 and refer to Figures A.1 to A.3 for the remaining standard compounds.

For each concentration, we found a rather small variation between the triplicates feature intensities (in general, two orders of magnitude smaller than the mean intensity) and thus a good reproducibility of both the LC-MS measurements and feature detection algorithm. The Pearson correlation coefficient between each compound's concentrations and corresponding feature intensities revealed an excellent linear relationship ($R \geq 0.98$ for all standard compounds). This result suggested that the relative quantification of analytes with our algorithm was reliable for a wide range of concentrations (in the presented example, about three orders of magnitude). In the case of the propionyl-L-carnitine-d3 standard, we could not detect any signals for the

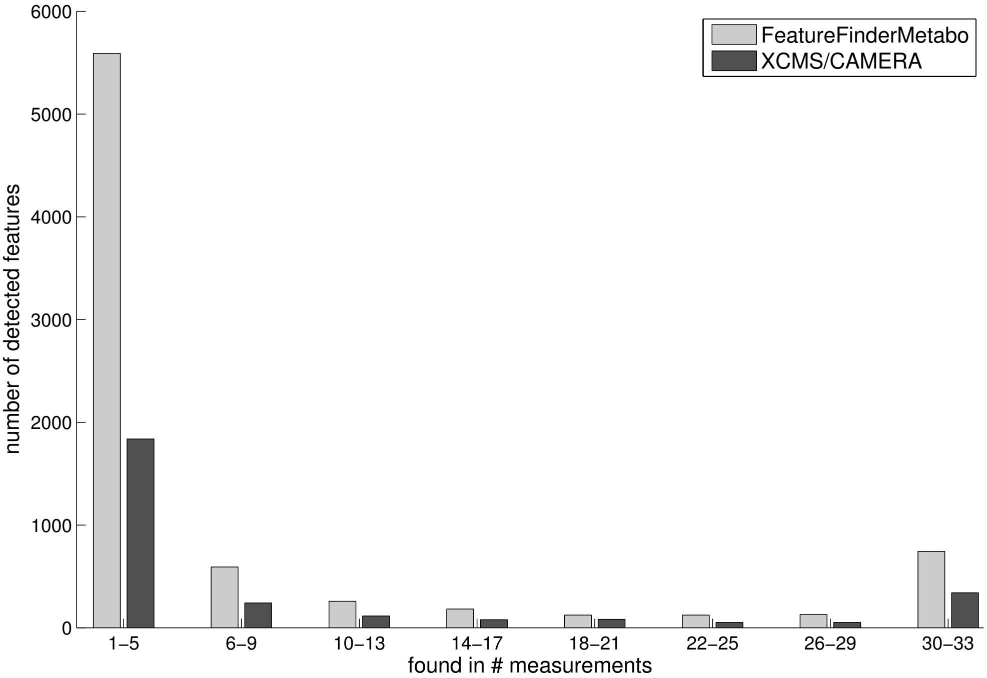


Figure 3.6: Numbers of reproducible features yielded by our algorithm and XCMS/CAMERA in the spike-in human plasma dataset. Each bin of the histogram plot represents a number of measurements a feature was reproducibly detected in. The bars show the absolute numbers of features that were detected with the respective reproducibility.

first six concentrations, neither by our algorithm nor by visual inspection of the raw data.

Apart from the standard compounds, the algorithm detected on average 2,020.0 ± 23.6 features per measurement whereof 351 were singly, 12 doubly, and six triply charged. The remaining features (on average 1,650) were singleton mass traces that exhibited no isotope pattern (e.g., due to low signal intensity) and thus did not allow determining the charge state. Since the ground truth set of real metabolites was unknown for the plasma samples, the distinction between real metabolite features and random features (e.g., noise, contaminants) was not possible. To resolve this issue, we determined the percentage of measurements in which each feature was detected. This detection rate is a measure of reproducibility and may serve as a rough guide to tell randomly occurring features apart from potential metabolite features. A bar plot shows the distribution of these detection rates (Figure 3.6).

We found a set of 1,168 features to be reproducible in at least 50 % of all measurements. We then compared this feature set to the features detected in each individual measurement and determined the size of the intersection. On average, we found 1,066.0 ± 3.6 reproducible features in each of the 33 measurements, that is, about 53 % with respect to the average feature number of 2,020. Querying these features against the Human Metabolome Database (HMDB) [128] with our `AccurateMassSearch` (AMS) tool (mass error tolerance of 5 ppm and default list of potential adducts), 768 of the accurate masses yielded a hit, i.e., about 74 % could be assigned to at least one potential metabolite ID.

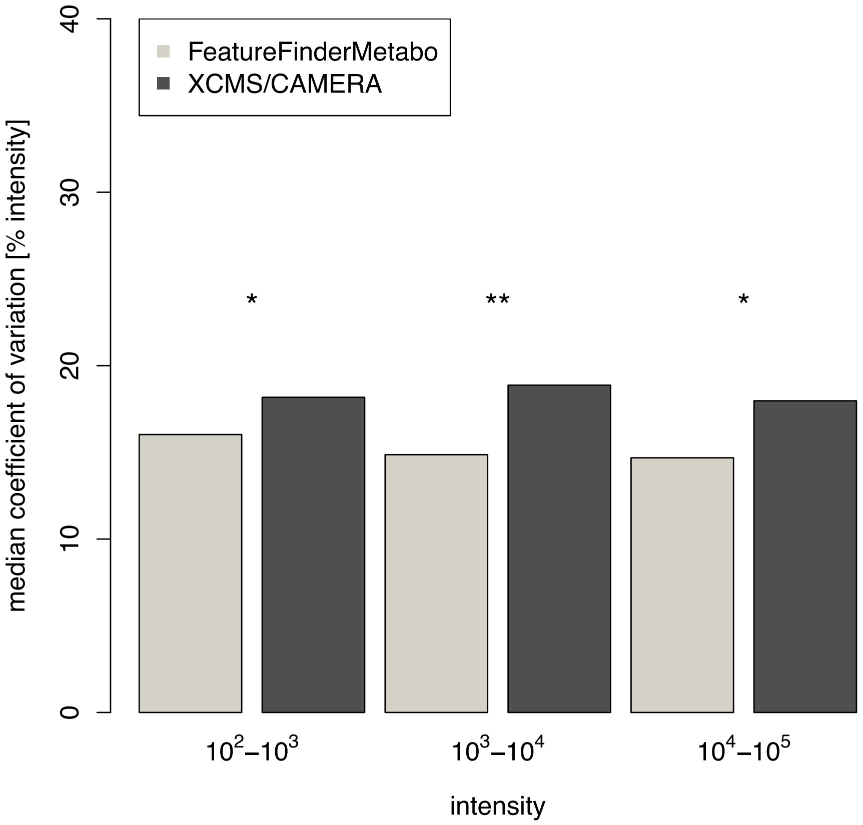


Figure 3.7: Intensity variation between features that were matched in at least half of the 33 spike-in human plasma measurements. For each matched feature, we computed the intensities' coefficient of variation (CV) over the respective measurements. All matched features were distributed to three bins with respect to the intensity's magnitude of order. Based on the intensity bins' underlying distributions, median CV values were determined and depicted in a bar plot. We conducted one-sided Wilcoxon tests to investigate if the observed difference of the median CV between the methods is significant. The results from the pairwise comparisons are marked by the significance level (asterisks). To allow for a direct comparison of FFM and XCMS/CAMERA feature intensities, we normalized all measurements by quantile normalization.

Finally, we considered the intensity variation of coinciding features between all measurements (features corresponding to the spiked-in standard compounds were excluded). We grouped these features into three intensity bins (i.e., magnitudes of order) to determine the dependence of feature quantification as a function of feature intensity (Figure 3.7).

In the first intensity bin (intensities between 10^2 and 10^3), we observed a median variation of about 16.0 %. The second intensity bin (intensities between 10^3 and 10^4) showed a median variation of 14.9 % that finally reached 14.7 % in the highest intensity bin (intensities between 10^4 and 10^5). These findings confirm a good reproducibility in terms of feature intensity when feature sets are compared between technical replicates even for low-intensity features.

3.4.2 Recall, Precision, and Robustness of the Algorithm

When simulating the dataset with parameters set to mimic experimental results most closely, we could observe excellent precision and recall of the features of 97 % and 96 %, respectively (Table 3.2). Unsurprisingly, increasing noise levels in the simulated data decreases both precision and recall of the method. In extreme noise settings, F-scores can thus drop to significantly lower values. The decline in performance primarily resulted from the loss in feature recall rate — detecting features reliably becomes more difficult as the signal-to-noise ratio decreases. In the following, we show detailed results from each of our simulation experiments.

First, we examined the influence of the noise_threshold_int parameter by adjusting it to each simulated detector noise level σ_{det}. For low detector noise levels, our algorithm achieves its maximum F-score of 0.96 (Figure A.4). Above $\sigma_{\text{det}} = 5$, the performance decreases moderately until reaching 0.62 on the highest noise level $\sigma_{\text{det}} = 100$. When filtering out mass spectrometric peaks with increasing values of this parameter, potential feature candidates are not considered which implies that the observed decline of the F-score results primarily from the decrease of the feature recall rate.

We studied the impact of under- and overestimating the mass error inherent to the data by considering two datasets each simulated with a specific mass error and varying configurations of the mass_error_ppm parameter (Figure A.5). For a simulated m/z error $\sigma_{m/z}$ of 10 ppm (Figure A.5, left), the algorithm achieves an F-score of 0.97 when the parameter is set accordingly. In comparison, we observe a similar performance for a simulated m/z error $\sigma_{m/z}$ of 40 ppm (Figure A.5, right). The performance plots in Figure A.5 suggest that the algorithm's performance is insensitive to moderate underestimation of the mass_error_ppm parameter. In case of overestimation, the performance remains stable on a high level comparable to the F-score yielded by the optimal settings due to the dynamic re-estimation of mass_error_ppm.

Analogously, we assessed the robustness of the chrom_fwhm parameter for elution profiles simulated with heavy distortion (Figure A.6). Based on the distribution of all simulated elution profiles, we estimated an average chromatographic peak width of eight seconds. The algorithm achieves its maximum performance of 0.9 with chrom_fwhm set to 8 s and remains stable for higher settings. For settings below eight seconds, the algorithm's performance decreases slightly until approximately four seconds and more steeply below this mark. This drop in performance is due to the loss of feature precision rate since very low settings of this parameter lead to insufficient smoothing of the distorted elution profiles and thus to their fragmentation. However, the performance plots reveal that the F-score remains insensitive to sharp changes within a rather broad margin of values lower than the optimal chrom_fwhm setting (between four and eight seconds). This suggests that the algorithm allows for a robust performance even if the parameter is moderately underestimated.

Method	Recall	Precision	F-score
FFM	0.96	0.97	0.97
XCMS/CAMERA	0.88	0.37	0.52

Table 3.2: Performance scores computed on the basis of the simulated plant metabolite dataset. The algorithmic parameters were set appropriately to the simulated characteristics of the data.

3.4.3 Comparison to Related Methods

By repeating the benchmarks described above with XCMS/CAMERA, we could compare the performance of our method to the one of the current leading algorithms in the field. Results of that comparison are summarized in Table 3.2. For each of the human plasma measurements, XCMS/CAMERA yielded 691.0 ± 7.2 features. We linked the feature sets of all 33 experiments and computed their detection rates (Figure 3.6). Features with a detection rate above 50 % were extracted, yielding a set of 544 reproducible features. For the comparison of XCMS/CAMERA with our algorithm in terms of performance, we considered the overlap between the reproducible feature sets. Both methods detected a set of 361 common features, whereas XCMS/CAMERA and our algorithm found 183 and 807 exclusive features, respectively. Thus, our algorithm detected a substantial proportion (66 %) of the XCMS/CAMERA feature set. At the same time, XCMS/CAMERA found only about 31 % of our algorithm's feature set. The high number of exclusive features indicates that our algorithm exhibits a higher sensitivity than XCMS/CAMERA. To verify the soundness of our exclusively detected feature set, we examined its intensity distribution and found the majority of feature intensities between 10^2 and 10^3. This suggests that the exclusive features stem mainly from a high sensitivity in the low-intensity range. Since we restricted this comparison to highly reproducible features, we assume our algorithm's exclusive features to originate from weakly concentrated analytes rather than spuriously detected artifacts, an observation that was confirmed by visual inspection of the datasets. With regard to intensity variation between replicate features, the comparison revealed that our algorithm exhibited slightly lower variations (14.7 to 16.0 %) than XCMS/CAMERA (18.0 to 18.9 %) across all intensity ranges (Figure 3.7). We observed that these differences in intensity variation were statistically significant between XCMS/CAMERA and our method, but not between different intensity ranges (based on a Wilcoxon test with $p < 0.05$).

We compared the benchmarking results of XCMS/CAMERA and our algorithm in terms of simulated mass error and chromatographic peak distortion. In both mass error simulations, our algorithm gave higher F-scores for the complete range of tested parameter settings than the XCMS/CAMERA counterpart (Figure A.5).

When setting the algorithms' mass error parameters according to the simulated error of 10 ppm (40 ppm), XCMS/CAMERA achieved an F-score of 0.52 (0.75) and our algorithm an F-score of 0.97 (0.95). Furthermore, the performance plots revealed differences in the robustness of the mass error parameter. While the parameter was insensitive to moderate under- and overestimation in our algorithm, in XCMS/CAMERA it was more susceptible to minor changes (Figure A.5)

3.4.4 Availability

The feature detection algorithm presented in this work was implemented and integrated in the open-source framework OpenMS (version 1.10) as the software tool `FeatureFinderMetabo` (FFM). The FFM tool can be invoked from the commandline or from `TOPPView`, OpenMS' graphical user interface for processing and analyzing MS data. Parameter settings are stored in XML-based configuration files (`*.ini`) that can be modified by using the `INIFileEditor` tool (Figure A.7).

As input file format, FFM accepts the mzML standard [57] and stores the results in the featureXML file format. FeatureXML files can be read and further processed by other OpenMS tools, e.g., `MapAlignmentPoseClustering` for rt correction or `FeatureLinkerUnlabeledQT` for linking coinciding features between multiple experiments. OpenMS offers functionality to build custom processing and analysis pipelines through the graphical workflow editor `TOPPAS` [118]. We employed `TOPPAS` to construct an example workflow for label-free metabolite quantification. This tool chain comprises feature detection, rt correction, feature linking, and the export to a tabular text file for subsequent statistical analyses. In order to allow for easy integration of downstream analysis methods such as statistical learning, we implemented this workflow in Konstanz information miner (KNIME), a powerful platform for information mining [129]. Both OpenMS 1.10 and workflows based on FFM are available online at `http://www.OpenMS.de`.

3.5 Discussion

The feature finding method introduced in this work, the OpenMS `FeatureFinderMetabo` (FFM), is a sensitive and reliable method for identifying metabolite features in LC-MS data. It is based on a sensitive mass trace detection and hypothesis-driven feature assembly. Model scoring with a hybrid mass deviation and isotope profile scoring results in accurately assembled features. An important component of the scoring function is the detection of likely metabolite profiles by statistical learning methods.

The algorithm can determine feature intensities of metabolites with high sensitivity and specificity (above 95 % on typical data). Even in complex backgrounds (human plasma) we found excellent linearity of the quantification based on repeated spike-in experiments. In a direct comparison, FFM was thus able to outperform established methods for LC-MS feature detection. Primarily in the low-intensity range, we are able to pick up signals with a poor signal-to-noise ratio. Thus, analyzing LC-MS data with FFM might reveal interesting new metabolites or at least expand the dynamic range of metabolomics analyses. Implementation of FFM as an OpenMS/TOPP tool ensures platform independence and convenient availability.

We designed the feature assembly algorithm to reconstruct solely isotope patterns. While it is a common strategy to also detect different adducts and charge states belonging to the same analyte and integrate these into a unique feature, we believe that this problem should be tackled in a separate processing step. Adducts of the same compound may show great variation in their ionization behavior and with that are difficult to detect accurately.

Future improvements of the algorithms will include a more advanced noise estimation model in order to increase the specificity of the algorithm. We are also exploring methods making use of multiple (replicate) measurements that collate information across multiple LC-MS runs in order to increase both the precision and the recall of the method.

CHAPTER 4

Metabolite Feature Identification

4.1 Introduction

It is a very common setting that untargeted metabolomics experiments yield hundreds to thousands of features that must first be identified before any biological interpretation can take place. While the knowledge of their identity is not a necessary requirement to take the untargeted approach, it is instrumental to generate more specific hypotheses and thus to reduce time and costs of targeted follow-up measurements. In this chapter, we present a comprehensive metabolite ID pipeline combining complementary methods.

We developed the `AccurateMassSearch` (AMS) tool to cope with a high volume of accurate mass search queries against the Human Metabolome Database (HMDB) efficiently and to annotate the feature maps coming from our `FeatureFinderMetabo` (FFM) tool. We chose the HMDB as our main data source since not only it covers the human metabolism comprehensively but also integrates metabolites from external sources (e.g., food or drugs) [27]. With more than 40,000 entries, it is one of the most complete metabolite resources publicly available. One of the main problems of accurate mass search is that the adduct behind an observed feature mass and hence its *neutral mass* (i.e., an analyte's mass in its uncharged state) is usually unknown. While it is common practice to focus on the most probable adducts such as the proton variants $[\mathrm{M}+\mathrm{H}]^{+}$ or $[\mathrm{M}-\mathrm{H}]^{-}$ only, many observed masses can be explained only if a broad spectrum of potential adducts is considered. To this end, the AMS tool was designed to be flexible in this regard, that is, the lists of potentially formed adducts can be modified freely depending on the experiment at hand. However, the more potential adducts are considered for the accurate mass search, the higher is the risk of false positive IDs. Exploiting the relative isotopic abundance

(RIA) obtained from high-resolution MS such as the recent Orbitrap devices were shown to improve the automatic annotation of metabolite signals [130]. Thus, we integrated an RIA filter with AMS in order to discard results with incompatible isotopic patterns. The final results are stored according to the mzTab standard, a column-based file format for reporting quantification and identification results [14].

The `MetaboliteSpectralMatcher` (MSM) tool integrates the basic functionality of spectral matching against metabolite fragment databases into the OpenMS framework. The success of spectral matching highly depends on the extent of metabolites covered in a spectral database. We chose the MassBank database as our main spectral resource [13], since it was one of the few public repositories that offered a good coverage of metabolite MS/MS spectra and that allowed its data to be retrieved for offline usage. While the default database is assembled from the MassBank repository, it is designed to be easily extended or replaced by in-house measured fragmentation data (e.g., spectra of authentic standards). We developed a specialized local data structure to hold the mass spectral information and to facilitate the efficient search of matching spectra. Analogously to the AMS tool, the results are stored in the mzTab format such that the coinciding IDs of both methods can be merged afterwards.

Depending on the mass error allowed and the number of hypothetical adducts considered during the AMS, the list of matching candidate empirical formulas and metabolite structures may grow dramatically. In order to reduce the number of false positive IDs, we studied the application of rt prediction models as an orthogonal filter criterion. The first hurdle to clear is often to assemble a representative training dataset preferably with a high number of metabolites that are annotated with structure and rt information. In case of untargeted experiments, however, there are usually either none or only a few identified metabolites to start with. We studied the applicability of rt models as orthogonal filters based on the RPLC-MS measurements as presented in Chapter 5. There, we experimentally validated a number of metabolites that could be exploited as a representative training dataset.

When the metabolite ID approaches are employed individually, they are often not powerful enough to guarantee unique IDs. In particular, the AMS tool may generate a high number of false positive candidate metabolites per feature mass since it focuses on mass alone and does not incorporate any chemical knowledge. To some extent, this is remedied by the RIA filtering approach, however, complete isotope patterns are rarely observed in case of low-abundance metabolites. For this reason, the integration of the AMS tool with the complementary information offered by the structure-based methods such as spectral matching or rt filtering seems particularly promising. We studied a strategy that combines the results gained from the AMS and MSM tools to a consensus in order to further increase the confidence of computational metabolite IDs.

4.2 Methods

We first introduce the experimental procedure to measure an RPLC-MS/MS dataset for benchmarking and validation purposes. Then, the ID methods and their algorithmic details are discussed individually. Finally, we describe a novel strategy which integrates the orthogonal information of the AMS and MSM tools to achieve more confident IDs.

4.2.1 RPLC-MS/MS Validation Dataset

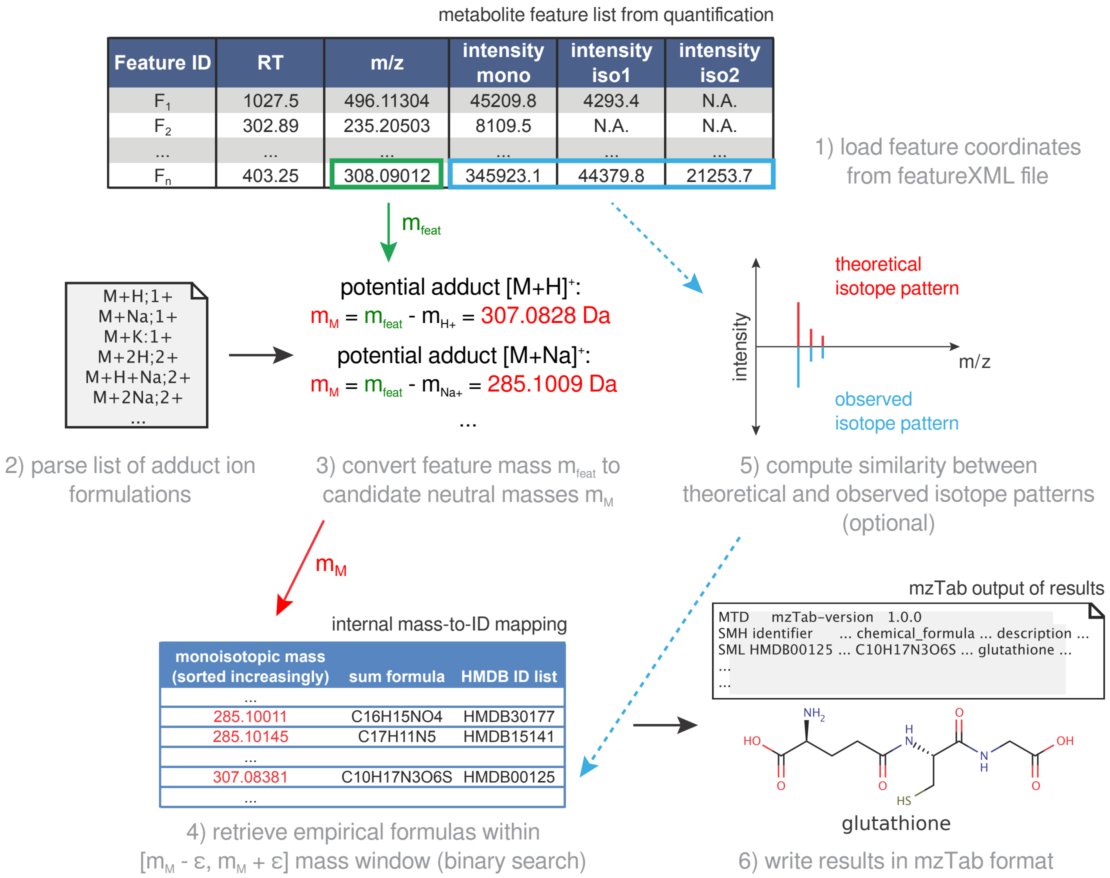
The following RPLC-MS/MS analysis was performed by Christina Ranninger (Division of Chemistry and Bioanalytics, University of Salzburg, Austria).

To validate our ID methods, a dataset was compiled from the authentic standards of all proteinogenic amino acids, ten metabolites associated with the TCA cycle, and 24 common metabolites (in total, 54 compounds); these were retrieved from our in-house chemicals storage (Table A.7). All standards were dissolved in varying amounts of substance in 30 % MeOH such that target concentrations between 10–100 µL were acquired.

The RPLC-MS/MS measurements were conducted with an Accela II HPLC system (Thermo Fisher Scientific, Bremen, Germany) coupled to an Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). The MS device was equipped with a heated electrospray ion source; the spray voltage was set to 3.0 kV in both the positive and negative ion mode. The chromatography was performed on a Hypersil Gold aQ column with 100 mm×2 mm in dimensions and 1.9 µm particle size (Thermo Fisher Scientific, MA, USA). In regard to mobile phases, we used water with 0.1 % formic acid (A) and acetonitrile with 0.1 % formic acid (B). The flow rate was 300 µl min^{-1}. The elution gradient was started with 100 % A for 1.5 min and then followed by a linear gradient to 100 % B during 6.5 min; 100 % B was held for 2 min. In the ending stage of the gradient, the column was re-equilibrated for 2 min at 0 % B. The total gradient time amounted to 13 min. During chromatography, the column temperature was maintained at 30 °C. For a single measurement, we injected 2.7 µl of standard mix solution.

The MS scan (100 to 1,000 m/z) resolution was set to 70,000 with an automatic gain control (AGC) target of $1 \cdot 10^6$ and a maximum injection time of 100 ms. The MS/MS experiment was performed with a resolution of 17,500 (AGC target was set to $1 \cdot 10^5$ and a maximum injection time of 50 ms). The isolation window was set to 1 Da and the applied normalized collision energy (NCE) was stepped (17.5, 25 and 32.5). The underfill ratio was set to 0.1 %, the apex trigger to 2–3 s, and a dynamic exclusion of 10 ms was chosen.

Both the positive and negative mode RPLC-MS/MS measurements were centroided and converted to the mzML standard with the ProteoWizard software suite (version 3.0.4243) [58]. We extracted the metabolite features from both datasets with the FFM tool (Chapter 3). For this purpose, we chose rather lenient parameter settings in order to capture the potentially



metabolite feature list from quantification

Feature ID	RT	m/z	intensity mono	intensity iso1	intensity iso2
F_1	1027.5	496.11304	45209.8	4293.4	N.A.
F_2	302.89	235.20503	8109.5	N.A.	N.A.
...	...	...	...	...	...
F_n	403.25	308.09012	345923.1	44379.8	21253.7

1) load feature coordinates from featureXML file
m_{feat}
M+H;1+
M+Na;1+
M+K;1+
M+2H;2+
M+H+Na;2+
M+2Na;2+
...
2) parse list of adduct ion formulations
potential adduct $[M+H]^+$:
$m_M = m_{feat} - m_{H+} = 307.0828$ Da
potential adduct $[M+Na]^+$:
$m_M = m_{feat} - m_{Na+} = 285.1009$ Da
...
3) convert feature mass m_{feat} to candidate neutral masses m_M
theoretical isotope pattern
intensity
m/z
observed isotope pattern
5) compute similarity between theoretical and observed isotope patterns (optional)
m_M
internal mass-to-ID mapping

monoisotopic mass (sorted increasingly)	sum formula	HMDB ID list
...		
285.10011	C16H15NO4	HMDB30177
285.10145	C17H11N5	HMDB15141
...		
307.08381	C10H17N3O6S	HMDB00125
...		

4) retrieve empirical formulas within $[m_M - \varepsilon, m_M + \varepsilon]$ mass window (binary search)
mzTab output of results

```
MTD    mzTab-version   1.0.0
SMH identifier     ... chemical_formula ... description ...
SML HMDB00125 ... C10H17N3O6S ... glutathione ...
...
...
```

Figure 4.1: Design of the search strategy as implemented by the AMS tool. (1) Feature coordinates (monoisotopic m/z, isotopic intensities) are parsed from the featureXML or consensusXML files. (2) A list of positive or negative adduct formulations are parsed and interpreted as calculation rules. (3) Based on these rules, the observed feature masses are converted to a list of putative neutral masses. (4) An internal table data structure is queried with each neutral mass to find matching empirical formulas within a user-defined mass error ϵ. (5) Optionally, isotope pattern filtering can be employed to filter out incompatible empirical formulas. (6) The ID results are exported in the tabular mzTab reporting format.

low-signal standards (Appendix A.1.2). The mzML and featureXML files constituted the input for the AMS and MSM ID runs described in Section 4.2.4.

4.2.2 Accurate Mass Search with Isotope Ratios Filtering

Based on a feature's observed mass m_{feat}, the AMS algorithm first computes potential neutral masses according to a predefined list of hypothetical adduct ions. This set of candidate neutral masses are then matched against the HMDB database with a user-defined mass error window to find compatible empirical formulas and to map these to actual metabolites. Although we focused on the data provided by the HMDB, the AMS algorithm can work with any other DB. The general strategy of the AMS algorithm is outlined in Figure 4.1.

The mapping between empirical formulas and matched molecule structures from the HMDB corresponds to the left-total relation $\mathrm{R}_{\mathrm{AMS}}$ between S, the set of empirical formulas present in the database, and H, the set of all unique metabolite structures, such that

$$\forall s \in S \; \exists h \in H : (s,h) \in \mathrm{R}_{\mathrm{AMS}} \tag{4.1}$$

holds (Figure 4.2). In order to cope with a high volume of feature mass queries efficiently (as it is the typical scenario with feature maps from untargeted metabolomics experiments), it was instrumental to construct a local data structure that modeled the $\mathrm{R}_{\mathrm{AMS}}$ mapping and offered offline searching functionality (i.e., stored in the OpenMS file hierarchy). We chose the HMDB database (version 3.5) as our default metabolite information resource [27]. First, we retrieved the data from the online repository comprising over 40,000 records. Each record is stored as an XML-based file containing basic information (e.g., monoisotopic mass, common name, simplified molecular input line entry specification (SMILES) or international chemical identifier (InChI) strings) and additional meta-data (e.g., physiological concentrations, related literature) about a specific metabolite. We implemented the Python script `parseHMDB30.py` (Appendix A.1.2) to extract those fields from each XML file that were essential for the $\mathrm{R}_{\mathrm{AMS}}$ mapping (empirical formulas, corresponding monoisotopic masses, and record identifiers). In Python, we modeled $\mathrm{R}_{\mathrm{AMS}}$ as a dictionary data structure with the empirical formula string as key and a list of matching record identifiers as value; whenever a previously encountered empirical formula was found in the dictionary, the record identifier was appended to the corresponding list. This dictionary was finally stored in the tabular text file `HMDBMappingFile.tsv` where each line corresponds to one empirical formula (Table A.4). Additionally, another dictionary was built as a look-up table to retrieve structural information about each HMDB entry and stored in the tabular text file `HMDB2StructMappingFile.tsv`. By loading the condensed information from these two files into map objects for fast lookup operations, our algorithm could assign structural information to observed feature masses in a fast and efficient manner.

Since the HMDB metabolites are reported as uncharged molecules by default, the observed feature mass must first be converted to its putative neutral mass. However, it is often not clear which specific adduct ion caused the mass observed and thus several putative adducts must be taken into account. The AMS algorithm was designed to allow for a flexible, user-definable list of adducts to consider. To this end, we implemented a simple parser that reads the pseudo-molecular formulation of an adduct ion and interprets it as a calculation rule in order to transform the adduct mass to the neutral mass. Before running the search algorithm, these adducts are read in from the `PositiveAdducts.tsv` or `NegativeAdducts.tsv` file, respectively. The default lists of potential adducts were adapted from the online resources of the FiehnLab [131], the same HMDB employs in their online search [27]. In comparison to the common writing convention of pseudo-molecular ions, we used a slightly abridged version without the square brackets and with the charge separated by the semicolon character. These changes simplified the parsing process since all operands (i.e., sub-molecules) could be extracted via string splitting operations and interpreted by the OpenMS library's `EmpiricalFormula`

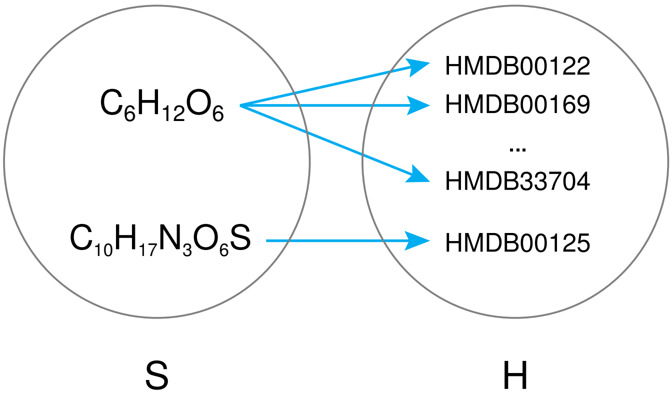


Figure 4.2: The mapping relation R_{AMS} modeled by the AMS tool. While each metabolite structure in H has a unique empirical formula in S, more than one distinct structure may be associated with the same empirical formula. For example, the empirical formula $C_6H_{12}O_6$ relates to several hexoses such as D-glucose (HMDB00122), D-mannose (HMDB00169), and D-galactose (HMDB33704). When the HMDB is queried for the empirical formula $C_{10}H_{17}N_3O_6S$, only glutathione (HMDB00125) is returned as a result.

objects to determine their monoisotopic masses (Figure A.9). Finally, the actual query mass was calculated by solving the adduct equation by the neutral mass m_M:

$$\begin{aligned} m_M &= m_{\text{feat}} - \frac{3}{2} \cdot m_{\text{H}_2\text{O}} - m_{\text{H}} + m_{e^-} \\ &= m_{\text{feat}} - 27.01585\,\text{Da} - 1.00783\,\text{Da} + 0.00055\,\text{Da} \end{aligned} \tag{4.2}$$

Based on this putative query mass, the AMS algorithm employs a binary search strategy to determine all empirical formulas of R_{AMS} that lie within $m_M \pm \epsilon$. The mass error ϵ depends on the mass accuracy provided by the MS device and can be set accordingly by a user parameter in the AMS tool configuration (either as absolute or relative error). The algorithm tracks the matching empirical formulas with a set of `AccurateMassSearchResult` objects (Figure A.8).

Even if the feature masses were measured with high mass accuracy, the AMS may yield several distinct empirical formulas for a given feature mass [65]. Furthermore, since we consider several putative adduct formulations simultaneously, there is a higher chance for finding distinct empirical formulas matching the same feature mass. In order to reduce the number of false positives, we applied an isotope pattern filter criterion as proposed in [65]. Given a feature with k isotopes and their intensities i_j, we compute the $k-1$ RIAs

$$r_j = \frac{i_j}{I} \quad \forall j = 1, 2, \ldots, k \tag{4.3}$$

where $I = \sum_0^{k-1} i_j$ is the total abundance of all isotopes. Analogously, for each matched empirical formula stored in the set of `AccurateMassSearchResult` objects, the ratios r'_j are computed from their theoretical isotope patterns. We define an empirical formula resulting from an AMS query as *valid* if the condition

$$|r'_j - r_j| < \sigma_{\text{RIA}} \quad \forall j = 1, 2, \ldots, k \tag{4.4}$$

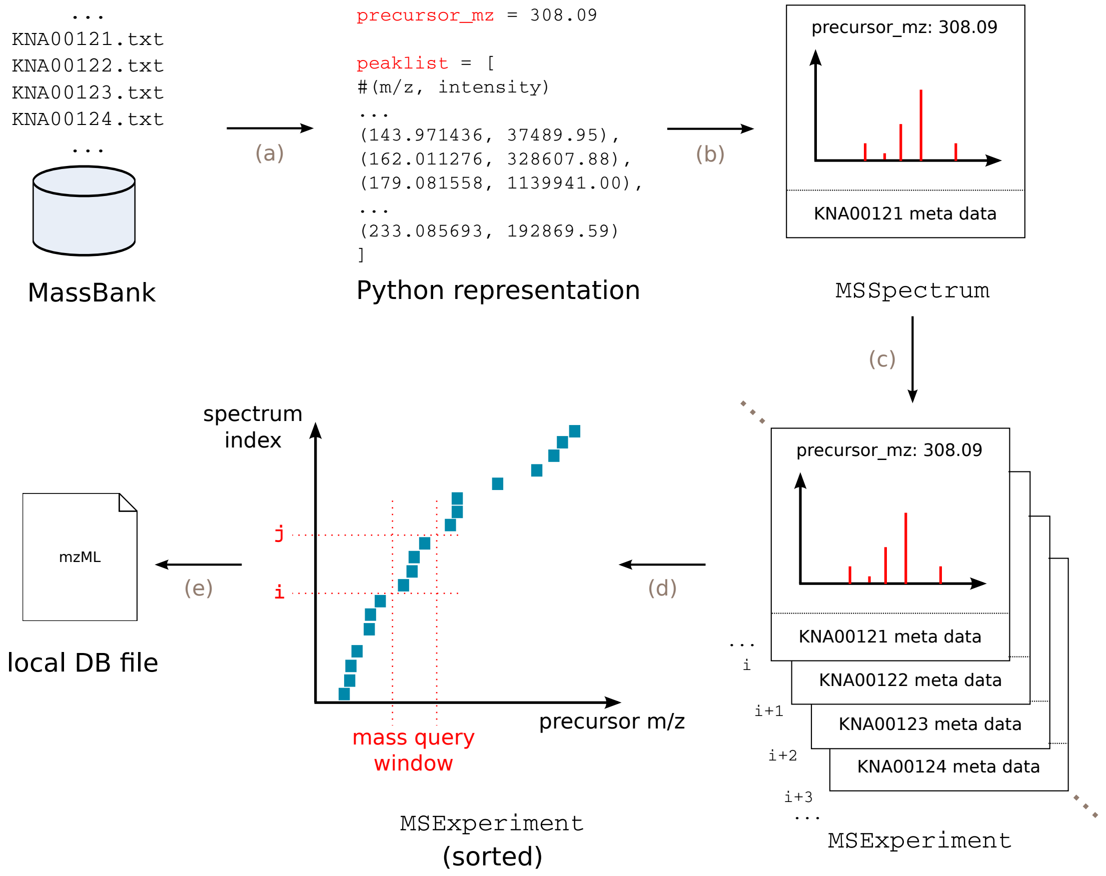
holds. The variable σ_{RIA} corresponds to the expected RMSE of relative isotope abundances and can be controlled by a user-definable parameter. In some cases, this RIA error is stated in the technical specifications of mass spectrometers and usually ranges from 2 to 5 % [66]. Sum formulas failing this condition are discarded from the set of AMS results while valid candidates are passed to the final report of results.

The AMS algorithm exports the putatively identified compounds according to the column-based mzTab standard specification [14]. Each potential metabolite structure (i.e., the unique record identifier) is written in an individual row; columns correspond to related information (e.g., common name, empirical formula) that are mandatory and are filled out with the data from the respective `AccurateMassSearchResult` object. The mzTab standard allows for the integration of non-standard information by appending optional columns. By this means, additional diagnostic information such as mass error and found adduct formula is exported to the output file. In order to facilitate the back-tracking of metabolite ID to the initially observed feature, each row is labeled by the feature index as stored in the original `featureXML` input file. For an example `mzTab` output file, we refer to Table A.5.

4.2.3 Spectral Search in Fragment Spectra Databases

Analogously to the requirements of our AMS tool to allow for efficient and offline database queries, our first step in implementing the MSM tool was to construct an efficient data structure. To this end, we retrieved the complete online data repository of MassBank, a total of 30,136 records [132]. Each record is stored as an individual text file containing the fragment peak list and additional information about how the spectrum was acquired (Table A.6). We implemented the Python script `parseMBFiles.py` to extract essential information such as the precursor mass, the product ion peak list, and metadata to classify the type of experiment (e.g., MS^n level, collision energy) from each record file (Appendix A.1.2). Both the peak and meta information were then transcribed by the script to OpenMS-based data structures to build a local spectra database stored in the mzML standard [57]. The general strategy of data extraction and processing is outlined in Figure 4.3.

As a result, the `parseMBFiles.py` script yielded the `MetaboliteSpectralDB.mzML` file that allowed us to load the spectral information from within the OpenMS file hierachy. During its initialization, our MSM tool loads this file into an `MSExperiment` data structure [134]. This map object completely resides in memory and thus facilitates fast access to individual spectra. Apart from that, since the spectra were sorted by ascending precursor m/z, it also offers to



```
precursor_mz = 308.09

peaklist = [
#(m/z, intensity)
...
(143.971436, 37489.95),
(162.011276, 328607.88),
(179.081558, 1139941.00),
...
(233.085693, 192869.59)
]
```

Figure 4.3: General data extraction strategy of the parseMassBank.py script. (a) The Python script parses for specific data fields such as metabolite identifier, precursor mass, or collision energy and extracts the peak list from each MassBank record. These data fields are filled into a temporary Python data object. (b) Using the functionality of the pyopenms package [133], the peak and meta information is filled from the temporary object into a MSSpectrum container [134]. (c) Each MSSpectrum object in turn is added to one MSExperiment container M, the OpenMS kernel data structure for storing LC-MS maps. A spectrum can be accessed efficiently via its container index. (d) The spectra contained in M are sorted by their precursor mass. This facilitates to quickly determine the set of fragment spectra (indices from i to j) that match a given precursor mass window. (e) The sorted data structure is stored to a local mzML file and thus allows for fast offline database queries.

restrict database queries only to those spectra with similar precursor masses. Given a specific query precursor mass m_p and error tolerance δm_p, the MSM algorithm determines lower and upper search boundaries, $m_p - \delta m_p$ and $m_p + \delta m_p$, and their corresponding spectra indices i and j within the map data structure via a twofold binary search (Figure 4.3). Subsequently, spectral matching is conducted only against the fragment spectra with indices ranging from i to j instead of processing the whole database.

To compute the spectral matching between an unidentified fragment spectrum S_{new} and the spectrum from a candidate metabolite S_k (where $k = i, i+1, \ldots, j$), we implemented a modified version of the probabilistic spectral similarity function proposed by Fenyö and Beavis [82]. While the original scoring function is geared towards the assessment of peptide fragmentation patterns and models the probability of observing a specific number of matching b and y ions [135] with a hypergeometric distribution, this was not applicable for the comparison of metabolite fragmentation patterns. Therefore, we relaxed the assumption of specific ion types such that all ions are treated the same and obtained the generalized score function

$$\text{HyperScoreModified}(S_{\text{new}}, S_k) = \log\left(\sum_{i=0}^{n} I_i \cdot I_i'\right) + \log n! \tag{4.5}$$

where I_i and I_i' correspond to the n matched product ion intensities from S_{new} and S_k, respectively. Here, the dot product describes the correlation between product ion intensities that match to each other within a specific mass error window. The second term models the probabilistic contribution to the score based on the hypergeometric distribution. The more product ions coincide between the two spectra in comparison, the less likely this spectral matching is due to chance alone and with that the higher the contribution to the overall score. Before a spectral matching score can be calculated, the algorithm must first determine the matching product ion peaks. For each observed product ion peak in S_{new} with mass m_p and mass error δm_p, the algorithm searches a peak in spectrum S_k that falls into the interval $[m_p - \delta m_p, m_p + \delta m_p]$ and is nearest to m_p. Since the spectra peaks are sorted by their m/z, these matchings can be computed efficiently by binary searches. The mass error δm_p allowed for product ions can be set by a user-definable parameter in the MSM tool's configuration.

All spectra S_k yielding a positive score when compared to S_{new} are recorded as putative metabolite candidates. These candidates are ranked by their score in descending order. The MSM tool offers the option to either report solely the best hit or a list of the top three hits. In case of multiple candidates, the algorithm assigns a group number that facilitates to track the putative IDs back to the original S_{new}. Finally, these results are exported in a fashion similar to the AMS tool as `mzTab` file (Table A.5).

In addition to the matching score, the algorithm allows the computation of an *E-value*, a measure of the statistical significance of the result [82]. In essence, this value states how far away it is from random spectral matches. Assuming that the positive scores follow a hypergeometric distribution, E-values can be computed with a linear function that was fitted to the steepest

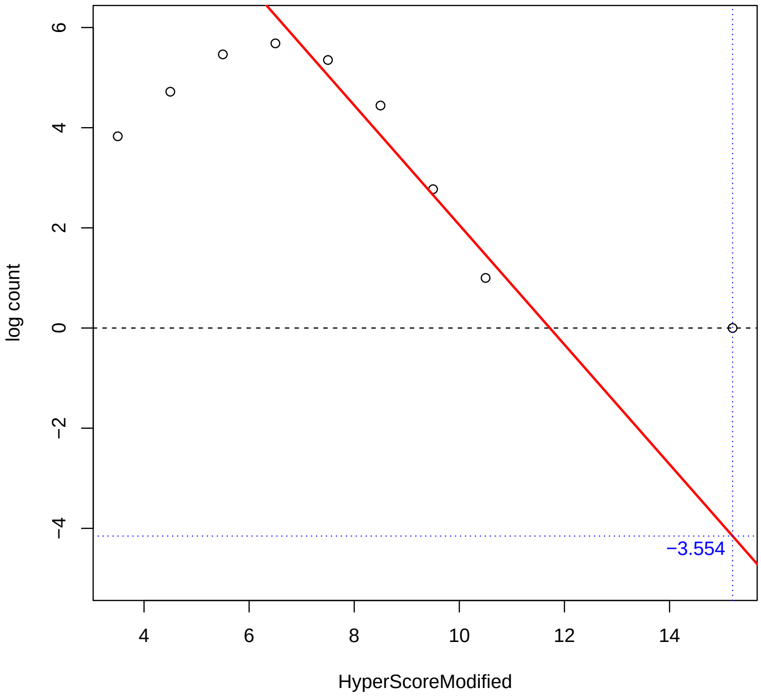


Figure 4.4: E-value computation based on an exemplary HyperScoreModified distribution. A regression line (red) is fitted to the data points that constitute the descending flank of the log histogram. The threshold for scores to be regarded as significant is determined by the regression line's zero crossing point on the x axis. In this example, the highest score has an E-value of $e^{-3.554}$ and thus it is very unlikely to occur by chance alone.

descent of the trailing edge (Figure 4.4). For each queried fragment spectrum, the algorithm first derives a histogram-based approximation of the scoring distribution by matching the query spectrum against the whole database and distributing the positive-scoring results to the histogram bins. Since there are multiple spectra recorded from the same compound that merely differ in their scores, the algorithm merely considers the best scoring spectrum for each metabolite structure. Based on the logarithmized counts, we determine the histogram bin with the maximum count and the subsequent bins that are characteristic for the distribution's trailing flank. The algorithm then fits a linear regression model to the corresponding data points. For each of the spectrum hits that are to be reported in the final `mzTab` output, an E-value is computed and added as a further optional column.

4.2.4 Exploiting the AMS and MSM Tools in a Combined Approach

The `mzML` and `featureXML` files generated by our RPLC-MS/MS experiments (Appendix A.1.2) constituted the input data for our MSM and AMS tools, respectively. We ran both tools separately in order to assess the individual performances first before evaluating them in combination. In practice, however, one could integrate both tools more tightly such that AMS acts as a prefilter step to reduce the number of potential precursor candidates and to speed up the following MSM run.

In case of AMS, we conducted an experiment with a 2 × 2 factorial design in which we studied the AMS tool's main parameters, the list of potential adducts considered and the mass error tolerance $|\delta m_{\text{AMS}}|$. With respect to the adduct configuration, we investigated two complementary scenarios. While we employed the complete adduct list as provided by default in the first scenario, we restricted this list to the most common $[\mathrm{M}+\mathrm{H}]^{+}$ and $[\mathrm{M}-\mathrm{H}]^{-}$ adducts in the second scenario (for adduct configurations, see Appendix A.1.2). In each case, AMS was run with the $|\delta m_{\text{AMS}}|$ parameter set to 5 ppm and 1 ppm. Furthermore, instead of reporting the best or top three hits per feature, the tool was configured to report all matching results. We investigated the frequencies of candidate IDs with respect to m/z in a heat map representation and studied the proportions of the different adduct types. The MSM tool was employed to match all MS/MS spectra recorded in the `mzML` files against the fragment database. For this purpose, we used the default parameter settings of our MSM (precursor and fragment ion mass error window set to 100 and 500 ppm, respectively). More details about the parameter settings of both tools can be found in their respective configuration files (Appendix A.1.2).

In order to determine all metabolites that were identified by both the AMS and MSM methods in an automated fashion, we implemented the Python script `mergeIDs.py` (Appendix A.1.2). In the beginning, the script computes the mapping of features (AMS) to precursors (MSM) that are closest to each other with respect to their rt and m/z coordinates. For each mapped feature and precursor pair, the script compared the ranked lists of IDs. In the case of AMS, the ranking was based on the increasing relative mass error between the observed feature mass and database hits; the MSM results were ranked by their decreasing spectral matching score (Equation (4.5)). We computed an *ID consensus* by matching the individual DB IDs and reported them together with their respective ranks in a tabular output file. If this matching was infeasible due to missing ID information, we performed a fuzzy string matching between the metabolites' common names instead (minimum similarity of 70 %). Although a structural matching would have been the most reliable option, this was infeasible for several MassBank records due to missing structure information. We confirmed the results by visually inspecting the MS data; in particular, we examined if the precursor rt and m/z coordinates coincided with a clear feature signal.

To assess the impact of the RIA filter on the resulting candidate numbers, we conducted several AMS runs with differing RIA tolerances (1, 2, 5, 10, 20, 50 and 100 %). Since this filter affects only features that exhibit an isotope pattern, we applied this analysis to a filtered `featureXML` file where all singleton mass traces were excluded. Similarly, the percentage of annotated features were determined under these parameter settings.

The validation of the combined approach was performed on the RPLC-MS/MS dataset introduced in Section 4.2.1.

4.2.5 Prediction of Metabolite Retention Times on RPLC-MS Columns

Based on the RPLC-MS dataset presented in Chapter 5, we assembled sets of 18 and 14 metabolites measured in positive and negative mode, respectively, with their rts covering the whole range of the elution gradient (Table A.8). We confirmed these metabolites experimentally with authentic standards or by MS/MS spectral matching. These initial datasets comprised the metabolites' identifiers, their m/z and rt as observed in the RPLC-MS measurement, and structural information (i.e., InChI strings). We used these data matrices as the starting point for the calculation of molecular descriptors.

We studied molecular descriptors that were simple and robust to compute from a compound's molecular structure and were predictive of their retention on RPLC columns. Intuitively, descriptors for a compound's hydrophobicity (expressed by the $\log P$ value [136]) were expected to correlate well with its retention behavior. We chose the XLogP descriptor [137] to estimate the $\log P$ for each of the training compounds. While the computations for our small dataset were manageable, it was instrumental to have a computational workflow that would automate these computations for a high number of unknown compounds to come. To this end, we implemented a solution in the KNIME software [129] that could process metabolite structures and their rts (e.g., from our AMS tool's `mzTab` output), compute their XLogP values, and finally export the descriptor matrix. The XLogP was computed with an open-source software package provided by the Pharmaceutical Data Exploration Laboratory (PaDEL) [138]. The PaDEL package implements a wide array of molecular descriptors and integrates this functionality into KNIME. The resulting descriptor matrix was stored as a tabular format file.

We constructed a second workflow that imports this descriptor matrix and builds a linear model to describe the relationship between the independent variable XLogP and dependent variable rt. The linear regression is computed with the `R` integration in KNIME (Figure 4.6). The rt prediction model was integrated in a *predictor node* that accepts new datasets (e.g., the `mzTab` file from our AMS tool) for future predictions. Moreover, we added a filter criterion to discard predictions that do not coincide with their observed rt values within the 95 % prediction interval, i.e., the corresponding rows were discarded from the input matrix. To further condense the number of overall putative IDs, we focused merely on simple adducts that occurred most likely (e.g, $[\mathrm{M+H}]^+$ or $[\mathrm{M-H}]^-$, see AMS adduct configuration files in Appendix A.1.2). Finally, the filtered IDs are stored in the same format as the input file.

4.3 Results

We first present the individual performances of our AMS and MSM tools. AMS detected most of the standards measured in the validation dataset. When used with default parameter settings, however, this came for the price of highly ambiguous IDs. This situation would curtail the usefulness of the AMS tool, in particular, for untargeted metabolomics analyses. To this end,

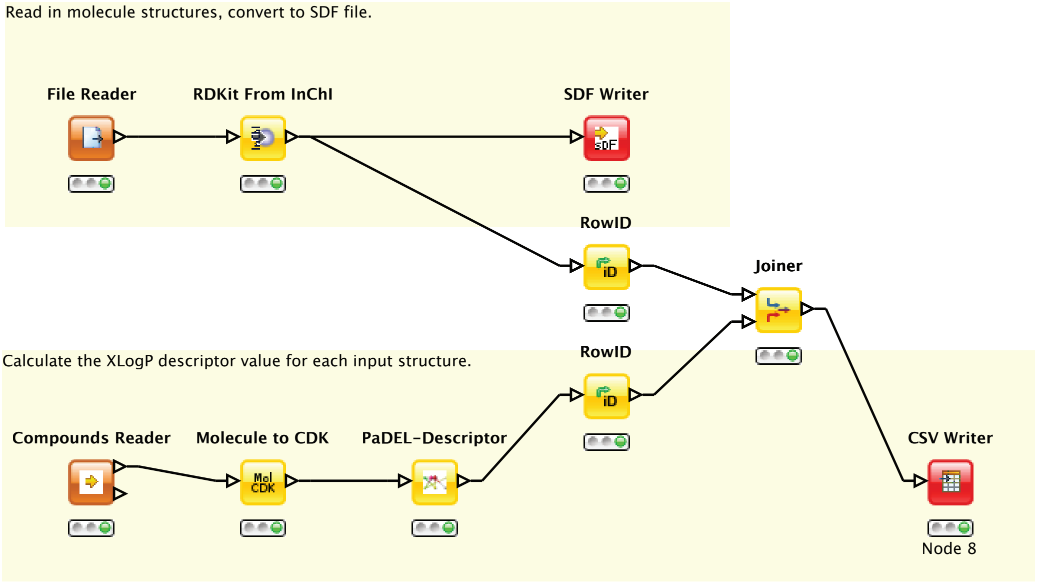


Figure 4.5: KNIME workflow for the computation of molecular descriptors. First, InChI structure identifiers are converted to the structure data format (SDF) format (upper workflow) with the RDKit package [139]. The compounds are then prepared by the chemistry development kit (CDK) node [140] (e.g., generation of a valid two-dimensional molecule structure by energy minimization). For each molecule, the XLogP value is predicted by the PaDEL node [138]. The descriptor matrix is written out in a tabular format and is ready for training an rt predictor or conducting new predictions.

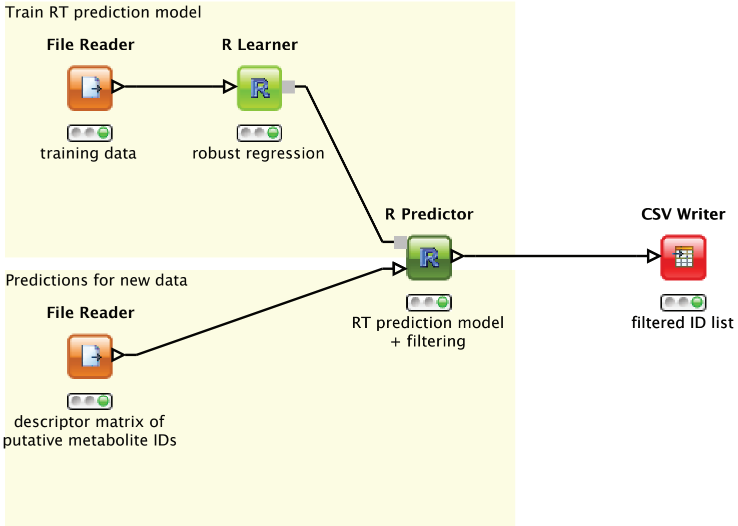


Figure 4.6: KNIME workflow for the training of an rt model and its application. Based on the descriptor matrix of compounds as training dataset, the rt model was built within the *R Learner* node either with ordinary or robust regression. This was feasible due to the simple integration of R code via KNIME's *R extensions* [8]. The resulting model is delivered to the *R Predictor* node and thus can be used for future predictions. Furthermore, additional logic can be added to the predictor node in order to filter out putative metabolite IDs with inadequate rts. The filtered list of putative metabolite IDs are exported again in a tabular file format.

	positive mode		negative mode	
#features	6,241		211	
w/isotopes	240		10	
$\lvert\delta m_{\text{AMS}}\rvert \leq 5\,\text{ppm}$	all adducts	$[\text{M}+\text{H}]^+$ only	all adducts	$[\text{M}-\text{H}]^-$ only
total #candidate IDs	80,757	13,258	3,306	461
#explained features	5,560 (89.1 %)	2,272 (36.4 %)	183 (86.7 %)	119 (56.4 %)
median #IDs/feature	8	3	12	3
$\lvert\delta m_{\text{AMS}}\rvert \leq 1\,\text{ppm}$	all adducts	$[\text{M}+\text{H}]^+$ only	all adducts	$[\text{M}-\text{H}]^-$ only
total #candidate IDs	22,773	6,166	44	5
#explained features	2,629 (42.1 %)	895 (14.3 %)	19 (9 %)	2 (1 %)
median #IDs/feature	3	4	1	2–3

Table 4.1: Overall statistics of AMS results from the validation dataset.

we performed several analyses to assess this ambiguity and to study the impact of filtering techniques such as RIA (Section 4.3.1). In contrast to AMS, MSM provided IDs that were much more evident due to the low number of false positive spectral matches (Section 4.3.2). However, it failed to detect some standard compounds partly due to the incompleteness of its spectral resource. We show the performance of the combined approach that could alleviate the tools' individual shortcomings (Section 4.3.3). Finally, we conclude this section by presenting rt models that were employed as filters of false positive AMS IDs in Chapter 5 (Section 4.3.4).

4.3.1 Individual Performance of AMS

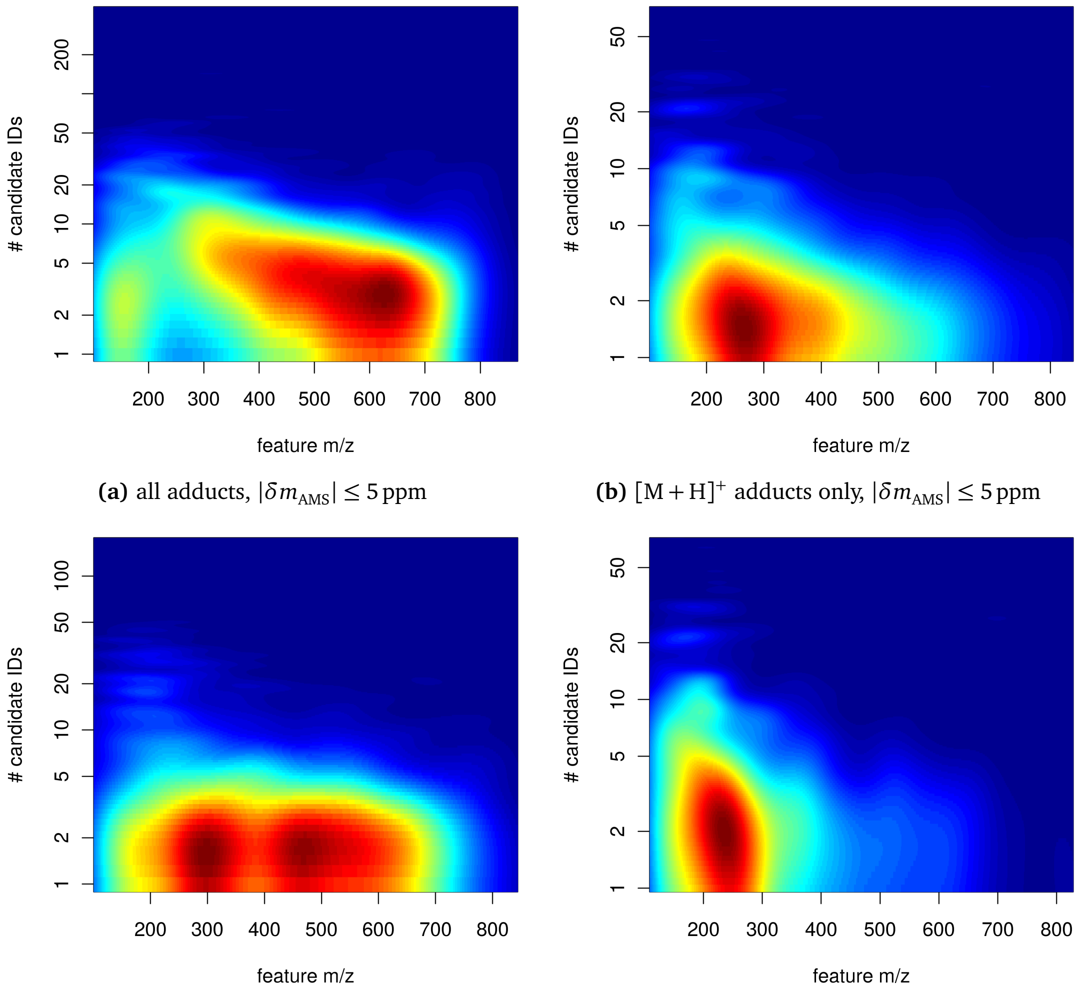
With an AMS run performed on our metabolite feature maps, we could identify 36 out of 54 compounds (66.7 %). From these, 25 compounds were ranked as best hits (i.e., showed the smallest difference from the observed feature mass). As expected from a common database query based on accurate mass alone, our IDs were overshadowed by numerous other false positive hits, rendering the interpretation of AMS results difficult. To some extent, such ambiguity is inherent to AMS due to isobaric metabolites. However, the huge number of candidate IDs compared to a much smaller number of metabolite features was striking (Table 4.1). For this reason, we further investigated how AMS' main parameter, the configuration of the adduct types to be queried and the mass error tolerance $\lvert\delta m_{\text{AMS}}\rvert$, affected the final number of candidate IDs.

As a rough estimate for the degree of ambiguity among the AMS results, we computed the median number of candidates proposed per metabolite feature (Table 4.1). As expected, the median number decreased considerably either when the list of adduct types was restricted or when the mass error tolerance was tightened. To study this behavior in more detail, we broke the candidate numbers per feature down with respect to m/z in a *heat map* visualization (Figure 4.7). When all adducts were considered with $\lvert\delta m_{\text{AMS}}\rvert \leq 5\,\text{ppm}$, we observed a remarkably broad and intense region extending from m/z 300 to 700 (Figure 4.7a), peaking at m/z 650 and two to five candidates per feature. That is, the bulk of candidate IDs were found for feature masses that lay well beyond the mass and rt range of the measured authentic standards (Table A.7). Under the assumption that the $[\mathrm{M+H}]^+$ adduct should occur most frequently, we would have expected a hot spot between m/z 100 to 200. At least, such a region was faintly recognizable in the very same heat map. The tightening of $|\delta m_{\mathrm{AMS}}|$ resulted in a massive drop of candidate ID numbers while a similar intense region as before was present (Figure 4.7c). However, its peak shifted to below m/z 500 and one to two candidates per feature. Additionally, a second prominent hot spot appeared peaking at m/z 300 and one to two candidates per feature that also included our metabolite features of interest. When the list of scanned adducts was restricted, this hot spot persisted and, with the tightest AMS configuration, became even more compact and pronounced (Figures 4.7b and 4.7d). We found that this hot spot coincided with a high-density region observed in the mass distribution of HMDB compounds (Figure A.10). Between m/z 100 to 500, the HMDB showed a high density of both unique and isobaric compounds (Figure A.11) that correlated well with the hot spot observed in Figure 4.7d. At the same time, the broad hot spot containing the high-mass features disappeared. With respect to the AMS settings considering the complete list of potential adducts (Figures 4.7a and 4.7c), we concluded that the hits observed in the higher mass range were due to less common adducts without an accompanying $[\mathrm{M + H}]^+$ sibling and thus most likely false positives.

Our heat map analysis confirmed that without the specific tailoring of the adduct list and with too lenient mass error tolerance settings, there was a high risk both to generate false positive annotations (based on spuriously detected features in the high m/z and rt ranges) and to introduce ambiguity into IDs that were otherwise clear, although the high proportion of seemingly explained features was tempting (Table 4.1). This was complemented by the observation that candidate IDs were spread evenly among several less likely adduct types instead of showing a prevalence of the most common $[\mathrm{M + H}]^+$ adduct (Figure A.12a). Upon the tightening of the $|\delta m_{\mathrm{AMS}}|$ parameter, the proportion of $[\mathrm{M + H}]^+$ adducts increased while the proportions of other less common adducts decreased (Figure A.12b).

We further assessed how the situation of high ambiguity could be remedied with the orthogonal RIA criterion. While a bulk proportion of the feature set consisted of single-trace signals as expected, a set of 225 features were detected with an isotope pattern that allowed for RIA filtering. With the filter option disabled, 82.7 % of these features were annotated while the list of their putative IDs had about 2,000 entries. After an RIA filtering with 5 % RMSE, 69.8 % were annotated and the number of candidates dropped to roughly 1,600 (25 %). The decrease in overall putative IDs was even more pronounced for tighter settings of the RIA filter (Figure 4.8). Apparently, for RIA error tolerance values that were typical for modern MS devices (i.e., in the range of 2–5 %), the filtering showed its strongest impact on ID numbers. More importantly, it had not erroneously affected any metabolite features that were expected from our list of authentic standards (Section 4.3.3).

With regard to the AMS runs on the negative mode dataset, we observed a similar situation



(a) all adducts, $|\delta m_{\mathrm{AMS}}| \leq 5\,\mathrm{ppm}$

(b) $[\mathrm{M}+\mathrm{H}]^{+}$ adducts only, $|\delta m_{\mathrm{AMS}}| \leq 5\,\mathrm{ppm}$

(c) all adducts, $|\delta m_{\mathrm{AMS}}| \leq 1\,\mathrm{ppm}$

(d) $[\mathrm{M}+\mathrm{H}]^{+}$ adducts only, $|\delta m_{\mathrm{AMS}}| \leq 1\,\mathrm{ppm}$

Figure 4.7: Heat map visualization of AMS results. The heat maps show the 2D distribution of feature mass with respect to the number of putative IDs per feature. Each heat map corresponds to one condition of a 2 × 2 factorial experiment where we varied the adducts list configuration and the mass error tolerance $|\delta m_{\mathrm{AMS}}|$ of our AMS tool.

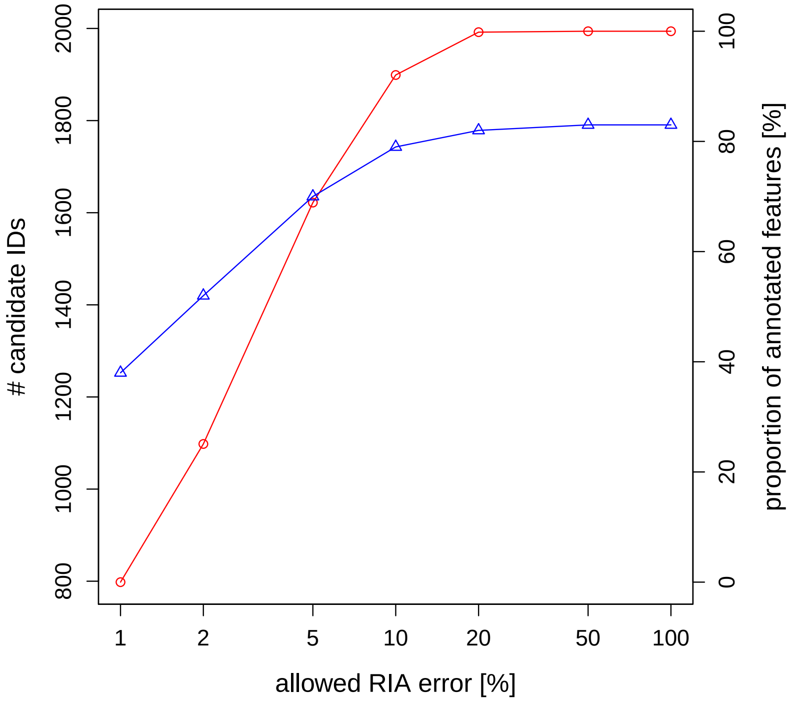


Figure 4.8: Impact of the RIA filtering on the number of candidate IDs (red line) and the percentage of annotated features (blue line). Candidate IDs were computed on the basis of 225 metabolite features with a detectable isotope pattern.

with highly ambiguous IDs (Table 4.1). The candidate IDs list ($|\delta m_{\mathrm{AMS}}| \leq 5\,\mathrm{ppm}$ setting) was dominated by $[\mathrm{M}+\mathrm{CH_3OH}-\mathrm{H}]^-$, $[\mathrm{M}-\mathrm{H}]^-$, and $[\mathrm{M}+\mathrm{C_2H_4OH}-\mathrm{H}]^-$ adducts (Figure A.13). Since there were too few cases for the RIA filtering to be applied to, the effect on ID numbers was expected to be negligible and thus was not considered further in our analysis.

4.3.2 Individual Performance of MSM

Given a total of 5,020 MS/MS spectra from the positive mode measurement, the MSM tool assigned 1,726 (34.4 %) to at least one positive-scoring match in the fragment DB. The median number of candidates per precursor mass was three. According to the E-value estimation, 783 (45.4 %) of these matchings showed a significant score. Discarding the results with insignificant scores, the median number of candidates per precursor dropped to two. We observed that the multiple matches to one precursor were often due to fragment spectra belonging to the same compound but yielding slightly different scores. This was expected since most compounds were recorded with varying collision energies or were deposited multiple times by different contributors.

MSM detected 28 out of 54 authentic standards (51.9 %). The majority of these were best-scoring hits (Table A.7). The only exception was in case of leucine (fourth rank) which scored slightly worse than its isomers isoleucine, norleucine, and allo-isoleucine. We computed the E-values of the MSM scores and found that 22 (78.6 %) matching scores were significant. Two spectral matches had a small positive E-value exponent (aspartic acid and NADH) while in case of the remaining four no E-value estimation was possible at all. The E-value computation can fail if there is not a sufficient number of spectra to estimate the score distribution (Section 4.2.3). Several compounds could not be identified since there was no detectable signal in the data. Aside from that, some unidentified compounds had no spectra in the DB.

4.3.3 Validation of the Combined Metabolite ID Approach

By merging the ID results from both the metabolite ID approaches, we observed consensus IDs in 27 out of the 54 cases (Table A.7). Among these, 19 (70.4 %) were ranked as the best-scoring hits by both the AMS and MSM tools. The majority of identified authentic standards came from the amino acids group (18 out of 20), only with glycine and alanine missing. The TCA and common compounds group showed the fewest consensus IDs (only 4/10 and 5/24 found, respectively).

In case of 17 standard compounds, neither AMS nor MSM could provide an ID. This agreed with the fact that no corresponding metabolite feature and thus no potential precursor ion selectable for fragmentation could be detected in the data. If we assume that these were *true negatives* and all compounds confirmed by only one method were regarded as *false* IDs, the accuracy of our combined ID approach would be 44 true assignments out of 54 (81.5 %). The compounds exclusively detected by AMS either had no spectra with matching precursor masses or no spectral information in MassBank at all. Some exclusive matches postulated less frequent adducts (e.g., $[\mathrm{M}+3\,\mathrm{CH_3CN}+2\,\mathrm{H}]^{2+}$) while precursor masses in MassBank usually originated from the most common $[\mathrm{M}+\mathrm{H}]^{+}$ and $[\mathrm{M}-\mathrm{H}]^{-}$ adducts. The ID accuracy of our combined ID approach could be further increased by confirming these exclusive AMS matches with other orthogonal criteria. However, neither RIA nor rt filtering were applicable in these cases.

4.3.4 Retention Time Prediction Models

We built two linear models $\mathrm{rt_{pos}}$ and $\mathrm{rt_{neg}}$, one for each positive and negative mode data (Figure 4.9). We found an excellent goodness of fit (leave-one-out (LOO) cross-validated R^2 of 0.93 for both models) along with high predictability power (F-statistic of 268.3 with p-value $< 2.0 \cdot 10^{-11}$ in case of $\mathrm{rt_{pos}}$ and 213.9 with p-value $< 5.2 \cdot 10^{-9}$ in case of $\mathrm{rt_{pos}}$). Based on the 95 % prediction intervals, the error margins of $\mathrm{rt_{pos}}$ ranged from ±2.5 to ±2.9 min and, in case of $\mathrm{rt_{neg}}$, from ±2.7 to ±3.2 min. We found that both models were of excellent quality and could be employed as rt filters for putative IDs.

We assessed the performance of our rt filters based on the datasets presented in Chapter 5. Other than described in Section 5.1.3, we did not restrict the AMS to proton-based adducts only but considered all potential adducts. From an effective rt filter we would expect that it decreases the median number of ambiguous IDs per metabolite feature. Furthermore, it

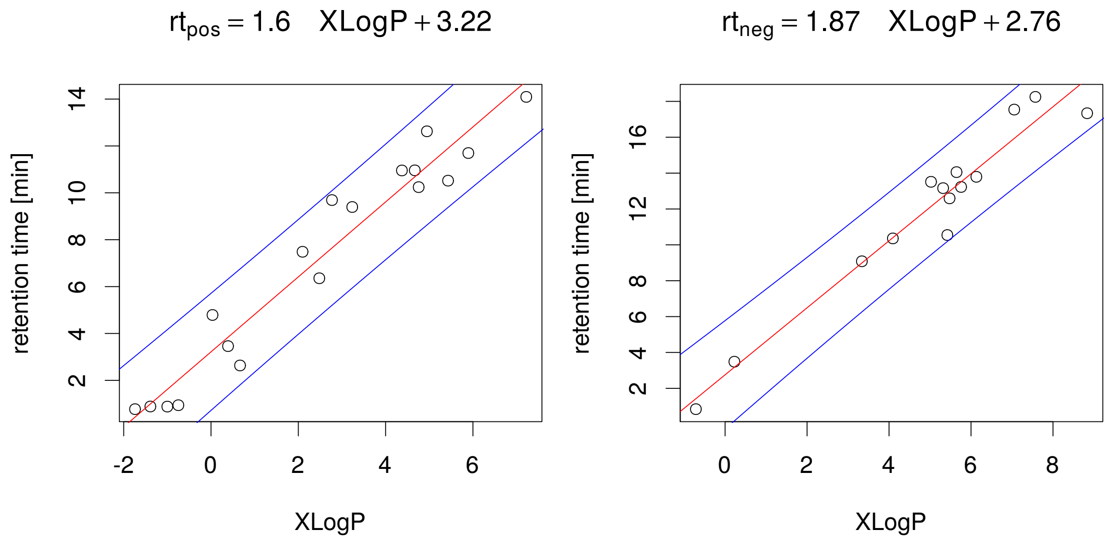


Figure 4.9: Retention time models employed as orthogonal filters for putative IDs. The linear models rt_{pos} and rt_{neg} (red lines) are depicted together with their respective 95 % prediction intervals (blue lines).

should also discard features that were spuriously annotated by chance. From our previous observations in Section 4.3.1, we know that this chance of spurious annotations increases the more potential adducts we consider in an AMS run. Based on the unfiltered positive mode ID list, AMS annotated 2,998 features with a median number of four putative IDs. After filtering these IDs with our rt_{pos} model, 1,170 features were annotated with the median number of two putative IDs. We further investigated this by superimposing the densities of ID numbers per feature and found that the rt filtering clearly resulted in an enrichment of less ambiguously annotated features (one to two putative IDs per feature) while reducing the number of features with high ambiguity (4 to 27 putative IDs per feature) (Figure 4.10). We observed similar results in case of the rt_{neg} model applied on the negative mode IDs.

4.4 Discussion

We presented a novel comprehensive metabolite ID strategy for untargeted LC-MS data that exploits several sources of orthogonal information to boost the confidence of putative IDs. Its performance was validated on the basis of a RPLC-MS/MS dataset comprising 54 authentic standards where we achieved an excellent ID accuracy of 81.5 % (IDs confirmed by both AMS and MSM). The loss in accuracy was mainly caused by MSM due to missing MS/MS spectra in MassBank. Another reason could be that the corresponding precursor ion was not selected during data-dependent acquisition (DDA). In these cases, AMS was more sensitive since it suggested the presence of nine other authentic standards. As we have shown in Section 4.3.1, however, these IDs must be treated with caution since they were not confirmed by another

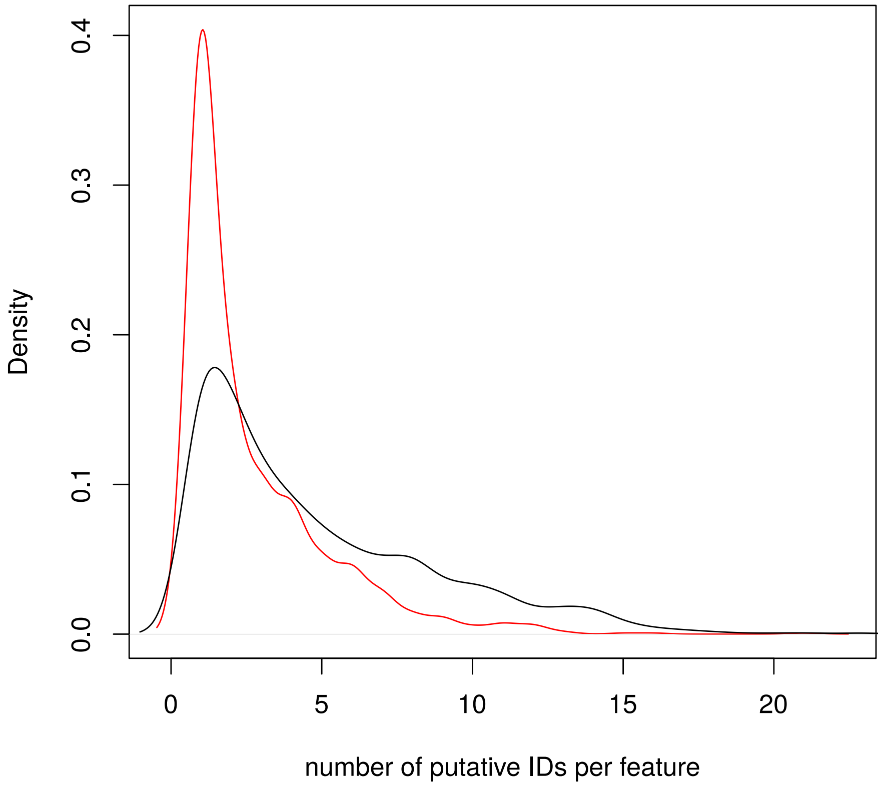


Figure 4.10: Densities of candidate ID numbers per feature before (black line) and after rt filtering (red line).

orthogonal criterion such as the RIA and rt filters. We could show in Sections 4.3.1 and 4.3.4 that these were powerful criteria to rule out false positive IDs and to reduce the high ambiguity of AMS results. However, neither the RIA nor the rt filters were applicable to the unconfirmed AMS results from our validation dataset. The underlying features were of low intensity and exhibited no isotopic pattern to be assessed by the RIA filter. With respect to rt filtering, the construction of representative prediction models was not feasible since most of the 27 concordantly confirmed compounds were strongly hydrophilic and thus covered only the first minutes of the elution gradient. We believe that the integration of these filter criteria would have further increased the confidence of our combined ID results.

We assessed the individual performances of AMS and MSM to learn about strengths and weaknesses when these tools were employed separately. AMS found evidence for most of the 54 authentic standards. However, it was very susceptible to lenient parameter settings, in which case the actual metabolite IDs were flooded by numerous false positives. Aside from isobaric metabolites (distinct compounds matching the same empirical formula), these incorrect IDs often stemmed from other adducts that fell into the same mass error window as the actual ID. This could be counteracted by stricter parameter settings, namely, a tighter mass error tolerance and a tailored list of the most common adducts (e.g., proton-based variants only). It has been shown before that mass accuracy alone does not guarantee a unique ID of a metabolite feature's empirical formula [65]. Setting the mass error window equal to the mass spectrometer's reported mass accuracy (e.g., 1 ppm in case of FTICR or Orbitraps) is a common practice, however, since it is an averaged value, AMS might miss several IDs with slightly higher mass errors. For instance, based on our validation dataset, we observed a mean mass error of (1.12 ± 1.33) ppm and thus a setting of 3–5 ppm was more adequate. An even stronger impact on the ambiguity of AMS ID results had the list of adducts to be queried for. We showed that relying on the default adduct lists (as employed by the HMDB online MS search) without tailoring them increased the number of putative IDs per feature considerably (Table 4.1). Aside from random hits due to the higher combinatorial possibilities, the co-occurrence of adduct rules such as $[M+H]^+$ and $[M+NH_4]^+$ may map different metabolites (e.g., aspartic and fumaric acid) to the same feature mass and thus introduce further ambiguity. This problem could be addressed by weighing all adducts by their likelihood of occurrence in ESI sources [68] or by the strict application of orthogonal criteria such as RIA and rt filtering if applicable. The RIA filter has been shown to be a powerful criterion to discard matches of false positive empirical formulas [65]. However, it might not be able to separate conflicting formulas if they differ only slightly and a typical RMSE of 2–5 % is assumed (e.g., as in case of aspartic and fumaric acid).

We showed in Section 4.3.4 that rt prediction models offer great potential to remedy the high ambiguity inherent to AMS results. The reason for this is that these models incorporate knowledge of the chemical structure into the otherwise solely mass-based AMS. We devised a novel computational workflow to automatically build rt models on the basis of IDs that were experimentally validated or concordantly confirmed by at least two orthogonal sources. The XLogP hydrophobicity value [137] of metabolites proved to be a robust molecular descriptor to build reliable rt prediction models for RPLC data. We could build linear regression models with as few as a dozen observations that showed excellent cross-validated performance. However, as a consequence of a large standard error, the prediction intervals were rather broad (±2.5 to ±2.9 min). These could be tightened by training the models on bigger datasets to allow for a better estimation of the standard error. Despite the large prediction intervals, our models were still very effective in filtering out false positive IDs and with that reducing the number of ambiguous IDs per feature considerably.

In contrast to AMS, MSM provided much clearer IDs. Most of the identified authentic standards were placed first with a significant matching score. With respect to the compounds that were detected exclusively by AMS, either MSM had no spectra with precursor masses compatible to the proposed adduct, or the compounds were not present in MassBank at all. Most of the MassBank spectra match to $[M+H]^+$ and $[M-H]^-$ adducts only and thus the chance that the more exotic adducts proposed by AMS (e.g., $[M+CH_3CN+H]^+$) are confirmed by MSM is low. The ID performance of MSM crucially depends on the completeness of the underlying spectral resource. It has been argued that the coverage of metabolites in public databases such as MassBank is not sufficient to allow for comprehensive ID of untargeted LC-MS/MS data [141]. The metabolite database METLIN provides more than 10,000 MS/MS spectra each measured in four different collision energies [142], however, is solely accessible via an online interface and thus can not be integrated into our MSM tool. The coverage is also essential for the computation of E-values to assess the significance of spectral matching scores. If there is not a sufficient number of spectra within the precursor mass error window, the accurate sampling of the hypergeometric scoring distribution is not feasible and with that no E-value can be estimated. This was the case for some of our compounds in our validation dataset (Table A.7). As a scoring function for spectral matching, we generalized the HyperScore [82] to MS/MS spectra of small molecules. In essence, it is a similarity function based on the dot product of fragment intensities matched between two MS/MS spectra and a probabilistic contribution. The dot product was widely used for matching fragmentation data from small molecules (e.g., by the National Institute of Standards and Technology (NIST) MS search [81]), however, novel scoring functions such as the X-Rank promise improved ID accuracy, in particular, when comparing MS/MS spectra from different instruments [84]. Regardless of whether another spectral DB or matching function is to be employed, these changes can be integrated quickly in the modular design of our MSM algorithm.

The initial results were very promising and clearly showed that our integrative metabolite ID approach holds great potential to address the common problem of identifying metabolites in untargeted LC-MS data. However, we must further assess our approach with other LC-MS/MS datasets to gain a more comprehensive picture of its performance. In particular, we would need data where all our orthogonal criteria could be combined. In the RPLC-MS/MS dataset presented here, the integration of the rt filter was not feasible. We believe that the overall performance of our combined approach can be further increased by improving its individual components. The interpretability of AMS results highly depends on the number of potential adducts that are included in the search. AMS in its present version treats each adduct as equally likely although they differ significantly in their ionization behavior in the ESI source and with that they exhibit quite distinct likelihoods of occurrence [68]. In order to improve on that, we plan to further investigate the default sets of potential positive and negative adducts as proposed by [131] and determine adduct likelihoods as weighing factors. Based on this, we would also like to integrate the likelihood for the co-occurrence of adducts (e.g., proton and sodium adducts). Furthermore, we want to implement a more fine-grained control over the various classes of metabolites to be searched for (i.e., different cell compartments, body fluids, exogenous sources such as foods or drugs). With regard to MSM, we plan to integrate MS/MS spectra data from other sources than MassBank. We would like to assess alternative scoring functions for spectral matches such as the X-Rank score that was adapted by the METLIN database [142].

CHAPTER 5

Analysis of Large-Scale Time-Resolved Metabolomics Data

A single liquid chromatography-mass spectrometry (LC-MS) measurement reflects the metabolic state of a biological system for a particular time point. Although such snapshots contain valuable information to study metabolic patterns responsible for specific phenotypes (e.g., disease states), the highly dynamic nature of the metabolome can not be captured by this approach. To elucidate the perturbation of metabolic pathways, e.g., during a physiological process or upon external stimuli, several timely successive snapshots must be measured in a time-series experiment. Temporally resolved experiments have been conducted to investigate the dynamics of plant, microbe, animal and human metabolomes [143, 144, 145, 146]. While time-series experiments, for example, with plants and microbes allow for a high time resolution, experiments on mammal and human samples must often rely on fewer time points due to limited material (e.g., blood, tissue biopsies). Furthermore, such experiments could carry a risk for animals and humans (e.g., severe injuries) and thus are strictly controlled by ethical restrictions. For instance, the changes in the human plasma metabolome during a single bout of exercise and recovery were investigated on the basis of only four time points [147].

Multivariate techniques such as principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) have become standard tools to investigate discriminant patterns of metabolites in metabolomics data (Section 2.3). However, both methods are not readily applicable to time-series data and thus need further adaption [148]. In contrast to these attribute-based multivariate methods, clustering techniques exploit the inherent similarity between observations to find recurring patterns in a dataset. The similarity between objects is computed through a distance measure (e.g., the Euclidean norm or Pearson correlation). A cluster of metabolite profiles might hint at a group of co-regulated metabolites or metabolic pathways or could assist in inferring the function of unexpected pathways under the studied experimental conditions. Thus, clustering has become a common data reduction technique to detect classes that exhibit a similar recurring pattern. However, the limitations of classical clustering approaches such as the hierarchical clustering [149] and k-means algorithm [150] fostered the development of novel clustering algorithms with parameters that were more intuitive to set and easier to interpret in biological experiments. The quality threshold (QT) clustering, initially developed for gene expression profile analysis [151], assembles clusters with a guaranteed quality, i.e., the distance between each data point in a cluster and its centroid must be lower than the predefined quality threshold. This allows filtering out clusters that are below a required minimum size and potentially stem from noise in the data. In the original work, the authors used jackknife correlation as a quality measure [151]; however, since the jackknife correlation is computationally expensive, variants based on the Euclidean norm or the Pearson correlation coefficient also exist. The QT criterion facilitates the interpretation of the resulting clusters tremendously. A remarkable feature of the QT clustering algorithm is its notion of noise. Data points not fulfilling the quality criterion of any potential cluster are removed from the dataset and marked as noise. This distinguishes the QT algorithm from classical approaches where each observation is strictly assigned to one of the clusters, even if its similarity to the cluster center is low. Although QT clustering is readily applicable to all kinds of temporally resolved data, it has been employed scarcely for metabolomics time-series experiments.

Physical exercise triggers massive changes in the human plasma metabolome that particularly reflect the energy metabolism of skeletal muscle. The detection of novel exercise-specific biomarkers promises to elucidate how physical activity benefits human health and to assess the risk of diseases if physical activity is lacking. Several time-series experiments have been conducted to investigate the physiological basis of physical exercise both with gas chromatography (GC)/LC-MS [147, 152, 153] and nuclear magnetic resonance (NMR) [154, 155] technologies. In other studies, the effect of carbohydrate and protein nutrition upon exercise [145] and the long-term changes induced by frequent physical exercise to the human metabolome [146, 156] were investigated. This increased interest in exercise-related metabolomics experiments revealed novel potential biomarkers, e.g., acylcarnitines that are assumed to support fat oxidation during moderate intensity exercise [147].

In this work, we present a novel strategy to investigate metabolomics time-course data collected during a single bout of exercise in male individuals. This data comprises hundreds to thousands of time profiles that are condensed to a few kinetic patterns that are characteristic of the metabolic changes induced by the physical exercise and the following recovery phase. To this end, we employ a stochastic variant of the QT clustering algorithm [23] that we extended by the short time-series (STS) distance [24]. The STS distance is particularly suited to short time series with non-equidistant time points. Additionally, we applied our computational approach based on `AccurateMassSearch` (AMS) and retention time (rt) prediction (Chapter 4) to identify key metabolites in strongly up- or down-regulated clusters. In an enrichment analysis, these metabolites were mapped to the Small Molecule Pathway Database (SMPDB) [157] to elucidate the metabolic pathways tentatively involved in exercise metabolism.

5.1 Methods

5.1.1 Ethic Statement, Subjects, and Study Design

The experimental design was published before [158] and shall be briefly recapitulated here. The study comprised nine healthy male individuals, however, one was excluded since the collected set of plasma samples was incomplete (i.e., missing time points). All participants were examined medically including common blood tests which showed no peculiar deviations. They did not engage in competitive sports but were considered as physically active (Table A.9). We informed them about potential risks and discomfort that may come with the experimental protocol, both orally and in writing. The protocol complied with the declaration of a Helsinki and was approved by the local ethical commission (H-D-2007-0127). We asked the participants to avoid exhausting exercises within the last 24 h before the experiments. After an overnight fast, we started our studies at 7 am. First, we inserted catheters to the resting leg's femoral artery and took blood samples to establish a baseline. Then, all participants did a one-legged knee extensor exercise for 2 h at 50 % of their maximum workload. We determined each individual maximum workload in pretrial runs [159], after the participants were familiar with carrying out the exercise. The exercise was performed on a modified Krogh ergometer. We took blood samples at three time points, both during the 2 h exercise (0, 60, 120 min) and the 3 h recovery stage (150, 180, 300 min). Finally, the catheter was removed when the recovery stage ended and the participants could eat and drink *ad libitum*.

5.1.2 Metabolite Signal Extraction

We extracted all metabolite ion signals from the LC-MS measurements with the Micromass MarkerLynx software (version 4.1, Waters, Manchester, UK). Metabolite ions stemming from the individual measurements were combined into one data matrix by aligning chromatographic peaks with similar mass (m/z error tolerance set to ±0.1) and retention time (rt error window set to ±0.2 min). The resulting data matrices (positive and negative mode, respectively) were used as the initial inputs for the subsequent data processing steps.

For each row of the initial data matrix, we used the concatenated rt and m/z information to uniquely label the corresponding metabolite signal (e.g., `10.95/344.2788`). In order to remove redundant isotope and adduct ion masses from our dataset, we employed our Python script `findFeatures.py` that merges metabolite signals compatible with respect to rt and m/z (Appendix A.1.3). For the detection of potential isotope masses, we scanned the list sorted by m/z to find signals similar in their rts and showing m/z differences that are approximately multiples of $1/z$ ($z = 1, 2, 3$). Additionally, we only grouped signals that showed a significant correlation between their intensities observed over all measurements (Pearson correlation of at least 0.6 with a significance level of $\alpha = 0.05$). If there were no compatible isotope masses, the reference m/z was added as a singleton to the final list of grouped signals. To find correlated proton, sodium, and potassium adducts in the positive mode dataset, the sorted mass list was scanned for signals similar in rt and showing m/z differences of 21.9819 (distance between a metabolite's proton and sodium adduct) and 37.9559 (distance between a metabolite's proton and potassium adduct). In case of the negative mode dataset, we searched for chloride (m/z difference of 35.9767) and bromide (m/z difference of 79.9262) adducts. For each of the merged ion groups, we selected the monoisotopic mass together with its observed intensities as the representative and discarded the corresponding isotopic and adduct masses. If the monoisotopic mass contained missing intensities, e.g. due to erroneous alignment, our `findFeatures.py` script exploited the correlated intensity information of isotopic/adduct ions to impute these. To account for the technical variation of metabolite signal intensities between the measurements, we applied quantile normalization as implemented in the `R` package `preprocessCore` [99].

5.1.3 Metabolite Identification

In order to putatively identify the ion masses extracted in the previous step, we employed our AMS ID workflow as presented in Section 4.2.2. The allowed m/z error window was set to 0.02. With respect to the potential adducts to be queried, we restricted AMS to search for the most likely proton-based adducts ($[\mathrm{M+H}]^{+}$, $[\mathrm{M+2\,H}]^{2+}$, $[\mathrm{M-H}]^{-}$, and $[\mathrm{M-2\,H}]^{2-}$) in order to keep the number of false positive IDs low. We accepted only IDs that were marked as endogenous plasma metabolites in the Human Metabolome Database (HMDB) and were cross-linked with pathway information from the SMPDB [157]. To further reduce the high number of false positive IDs, we built two linear rt prediction models (for positive and negative mode, respectively) on the basis of 32 experimentally confirmed metabolites (Section 4.2.5). We employed these rt models as orthogonal filter criteria to remove highly unlikely results from the AMS. If the difference between a putative ID's predicted and the ion mass' actually observed rt were not within the model's 95 % prediction interval, the putative ID was discarded from the list. The filtered ID lists were stored as R matrices to facilitate the quick lookup of IDs with an ion's unique label. We implemented the `R` script `clusterIDmapping.R` (Appendix A.1.3) that automates the matching of time profiles to our putative ID lists.

5.1.4 Cluster Analysis

The following steps were conducted within the R environment [8]. We rearranged the condensed data matrix in such a manner that all experiments corresponding to the same time point ($t_0, t_1, \dots, t_5$) from each individual were concatenated in the same column. To avoid ambiguity between the time profiles stemming from different individuals, the row labels were extended with their source (e.g., indB, indC, ..., indI)*. For each time profile, we computed $\log_2$ fold changes with respect to its median intensity. Based on these fold changes, we discarded time profiles that either showed no significant fold changes at all or had significant fold changes in both exercise and recovery stages (i.e., $\log_2$ fold change outside of the $[-1, 1]$ interval). To facilitate the comparison of time profiles with respect to their biological response range, each time profile's absolute intensities were transformed with range scaling [160].

The filtered time-series dataset was then clustered by employing the R package *flexclust* (version 1.3-4) [23]. This package implements a stochastic variant of the QT clustering algorithm that considers a small randomized subset of data points as initial cluster centers and thus allows for a considerable reduction in computation time. The Euclidean distance is offered as the default option; however, this distance function has severe shortcomings when considering time series with non-equidistant time points. To resolve this issue, we implemented the STS distance function [24] and integrated it into the flexclust R package (Appendix A.1.3). Given two time profiles $f = (f_0, f_1, \dots, f_5)$ and $g = (g_0, g_1, \dots, g_5)$, the STS distance is defined as

$$d_{STS}^2(f,g) = \sum_{k=0}^{n_t-1} \left(\frac{f_{k+1} - f_k}{t_{k+1} - t_k} - \frac{g_{k+1} - g_k}{t_{k+1} - t_k} \right)^2 \tag{5.1}$$

The STS distance considers the piece-wise slope differences between two time profiles and thus implicitly takes varying time intervals into account. In the context of our RPLC-MS experiments, the time points were at $t = (0, 60, 120, 150, 180, 300)$ min. Contrary to common distance measures such as the Euclidean metric, differences in short time intervals are penalized more than in longer intervals. Applied to metabolomics time-profile data, this means that we put emphasis on metabolic processes occurring in short time intervals, whereas in long time intervals, we take the time profile's greater variability into account.

We configured the qtclust function's parameters as follows: The clusters' radius parameter was set to 0.5. The minimum expected number of time profiles in a cluster min.size and the size of the randomized sample ntry were set via the control parameter to 20 and 3,000, respectively. The STS distance function was selected via the STS parameter of qtclust. To extract and visualize the individual clusters from the resulting qtclust object, we implemented the R function extractClusters. This function generates plots and box plots (intensity distribution for each time point) for a subset of clusters that accounts for 80 % of the dataset's time profiles. Additionally, for each individual cluster, the time profiles are extracted in a separate tabular file to allow for metabolite ID and pathway enrichment analysis in the subsequent analysis. Each cluster was designated by qtCl followed by its index within the qtclust object (e.g., qtCl3). We assembled the presented functionality in the R script qtClustPipeline.R (Appendix A.1.3).

*Originally, the set of labels referred to *nine* individuals starting with "indA". Due to missing data, we excluded "indA" but kept the labels "indB" to "indI" for the sake of simplicity.

5.1.5 Pathway Enrichment Analysis

The enrichment analysis enabled us to further condense the list of putative metabolite IDs. The rationale behind this was that the true metabolite IDs should show up significantly enriched in specific pathways, whereas false positive IDs are spread randomly over all pathways without any significant enrichment. Based on our metabolite ID table, we assigned up to three ID candidates to each clustered time profile. We extracted all pathway information contained in the SMPDB to facilitate the mapping of HMDB to SMPDB IDs (that is, metabolites to their associated metabolic pathways). We implemented the Python script `hmdbPathwayMapper.py` to perform the following steps (Appendix A.1.3). First, we determined an individual cluster's distribution of HMDB IDs over the metabolic pathways. Multiply occurring HMDB IDs were counted once. Then, for each pathway, we applied a one-sided Fisher's exact test to investigate if the ratio between matched and total number of HMDB IDs was significantly greater or due to chance. We sorted the enriched pathways by their significance and reported results with *p*-values below 0.05.

5.1.6 UPLC-qTOF-MS Analysis

According to the protocol published in [161], the protein content of plasma samples were precipitated with acetonitrile and then run to dryness. Before the analysis, the samples were reconstituted in 100 µL of a 4:1 mixture of acetonitrile and water. These were chromatographically separated with a 2.1×100 mm ACQUITY^{TM} 1.7 µm C8 column (Waters, Dublin, Ireland) at 40 °C operated in an ACQUITY-UPLC system (Waters Corp, Milford, USA). The elution gradient was programmed to start with 95 % A (A = 0.1 % formic acid in water) for 0.5 min, then to linearly increase to 100 % B (B = acetonitrile) during the next 24.5 min and to hold for 4 min, and finally to switch back to 95 % A. The column's flow rate was set to 0.35 mL min^{-1}. The UPLC system was coupled to the electrospray ionization (ESI) source of a QTOF mass spectrometer (Micromass, Manchester, UK). MS measurements were performed in full scan mode from m/z 80 to 1,000. The source temperature was set to 120 °C with a cone gas flow of 50 L h^{-1}. The desolvation gas temperature was at 300 °C and the gas flow at 500 L h^{-1}. In case of the positive ion mode, the capillary voltage was set to 3,000 V. For negative ion mode, a capillary voltage of 2,600 V was configured. For both modes, the cone voltage was at 30 V.

5.2 Results

We extracted 4,770 and 689 ion masses from the positive mode and negative mode measurements, respectively. These initial matrices were condensed to 4,201 and 494 masses by filtering out redundant isotope and adduct masses with our `findFeatures.py` script. We applied our adaption of the QT clustering algorithm to the filtered and combined matrix of 8,877 time profiles (7,854 profiles from positive and 1,023 from negative mode). This filtered dataset

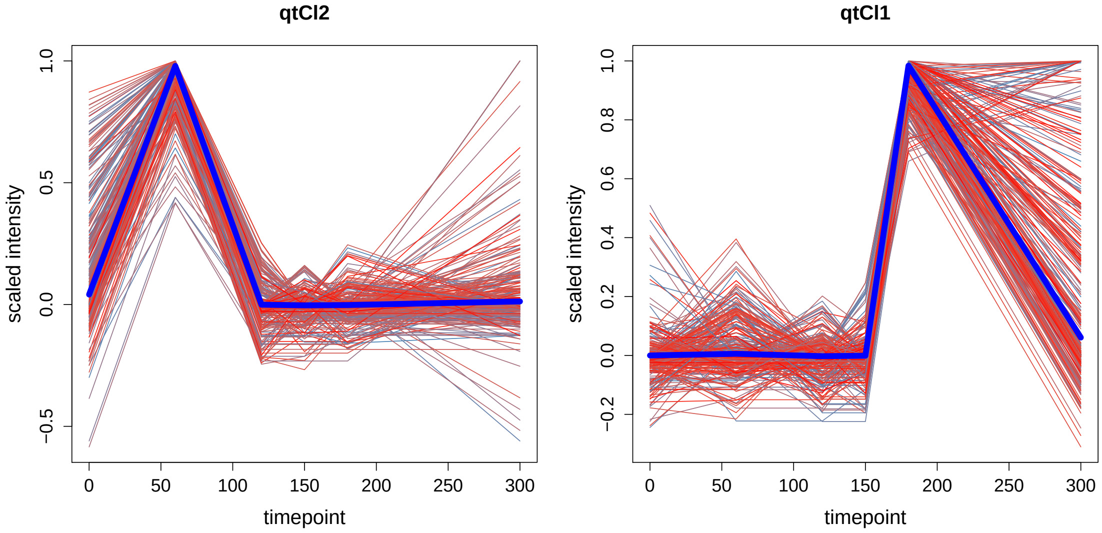


Figure 5.1: Clusters representative for metabolite perturbations in the exercise and recovery phases. According to the QT clustering results, these were the two largest clusters (1,261 and 1,060 time profiles in `qtCl1` and `qtCl2`, respectively). Both showed a clear increase in metabolite levels during the respective phases.

comprised 2,925 of the original ion masses from the positive and 353 masses from the negative mode measurements. The clustering step condensed the data to 25 distinct clusters with sizes ranging from 54 to 1,261 profiles. About 26 % of the time profiles were not assigned to any cluster. By visual inspection, we assigned 13 clusters to the exercise (Figures A.14 and A.15) and 12 clusters to the recovery stage (Figures A.16 and A.17). From these, we picked the two largest clusters `qtCl2` and `qtCl1` that were also the most striking representatives for the exercise and recovery stages, respectively (Figure 5.1).

We experimentally confirmed the identity of 32 metabolites, 18 from positive and 14 from negative mode data (Table A.8). These metabolites established the basis for two rt models of excellent quality which then were applied on the AMS results as orthogonal filters (Section 4.2.5). As a result, we could reduce the ambiguity introduced by false positive AMS results significantly and increase the confidence of the remaining IDs. In case of the `qtCl2` and `qtCl1` clusters, we could annotate 78 and 132 time profiles, respectively.

We performed a pathway enrichment analysis on these putative ID lists and found characteristic sets of pathways perturbed during the exercise and recovery stages (Table 5.1). Most of the proposed pathways were related to amino acid metabolism, a prominent subject of exercise-related research (e.g., amino acid supplementation to increase muscle protein gain [162]). Furthermore, we found pathways related to the oxidation of fatty acids, homocysteine degradation, betaine, taurine/hypotaurine, and vitamin B6 metabolism. Again, these metabolic pathways were frequently studied with respect to exercise [163, 164, 165, 166, 167]. Most of enriched pathways were exclusive to the exercise or recovery stage. The only exception was the catecholamine biosynthesis pathway that was common to both stages, although it was more significant in the recovery stage. The level of catecholamines (i.e., adrenaline, noradrenaline, and dopamine) were shown to be increased during exercise [168]. They are synthesized from the amino acids tyrosine and phenylalanine whose pathways were also enriched, however, exclusively in the recovery stage. The perturbation in tyrosine and phenylalanine levels might be explained by protein degradation in the exercising muscle [169] and could fuel the biosynthesis of catecholamines during recovery.

SMPDB pathway	exercise	recovery
glycine and serine metabolism	**	
glutamate metabolism	**	
betaine metabolism	**	
methionine metabolism	*	
oxidation of branched chain fatty acids	*	
beta oxidation of very long chain fatty acids	*	
homocysteine degradation	*	
arginine and proline metabolism	*	
taurine and hypotaurine metabolism	*	
catecholamine biosynthesis	*	**
tyrosine metabolism		**
vitamin B6 metabolism		*
phenylalanine and tyrosine metabolism		*

Table 5.1: Significantly enriched metabolic pathways in the exercise and recovery stages. The number of asterisks designates the significance of enrichment (* if $p < 0.05$, ** if $p < 0.01$).

5.3 Discussion

With our computational pipeline, we condensed thousands of time profiles to a manageable number of clusters that were characteristic for either the exercise or recovery stages. Our strategy was based on the stochastic QT clustering algorithm [23] which we adapted and extended by the STS distance [24]. Our application of the STS distance was clearly geared towards a common problem in metabolomics time-series experiments. Usually, there are only few time points due to cost and limited sample material. Furthermore, time points are not always equidistant and thus the clustering algorithm must deal with profiles that are highly heterogeneous in their time intervals with respect to biological variation. For instance, we would expect higher variability in metabolite levels in the last time interval (180 to 300 min) than in the shorter intervals before (see high scattering of the last time point in Figure 5.1). In such settings, we found that the STS distance was more adequate in matching metabolites with related kinetics than the classical Euclidean distance or Pearson correlation.

The two largest clusters `qtCl2` and `qtCl1` showed a clear peak in metabolite levels during the exercise and recovery phases, respectively. Since great proportions of the 8,877 time profiles and also the highest number of filtered putative IDs were distributed to these two clusters, we regarded them as strongly correlating with the metabolic processes taking place during the exercise and recovery phases. This assumption was further confirmed by assessing these clusters' metabolite ID in a pathway enrichment analysis. Here, we found metabolic pathways that were frequently investigated in the context of physical exercise such as amino acid or fatty acid metabolism [162, 163]. In summary, we found that our novel computational pipeline was capable of condensing untargeted and timely-resolved LC-MS data to a few prominent metabolite patterns that could be correlated with meaningful functional information from the SMPDB.

A computational strategy to investigate the response to physical stimuli has been published before [152]. It employed a scoring function that allowed for the prioritization of metabolites showing significant changes throughout the exercise session. Furthermore, it allowed to infer the network of pathway reactions. The authors proposed a list of exercise-associated metabolic pathways that comprised the metabolism of several amino acids as in our case. However, since the method was validated with a targeted MS/MS experiment and thus depends on experimentally confirmed metabolite IDs, it is not clear if it is readily applicable to untargeted metabolomics experiments as well. This sets our method apart as a screening utility that takes all measured analytes into account and assists in the setup of targeted follow-up experiments.

While the Kyoto Encyclopedia of Genes and Genomes (KEGG) resource is widely regarded as the reference database when dealing with metabolic pathways, we decided to exploit the information assembled in the SMPDB instead. The SMPDB (version 2.0) covers more metabolic pathways than KEGG, in particular, with regard to metabolite signaling and physiological action pathways [157]. All SMPDB pathways are crosslinked with metabolites contained in the HMDB and thus to a rich content of physiological information and literature. This additional information assists in the physiological interpretation of the metabolites that were found by our screening approach. Conversely, we can directly map IDs resulting from our AMS tool back to SMPDB pathways (e.g., for the enrichment analysis).

In untargeted metabolomics workflows, the validation of metabolite IDs constitutes the most time-consuming step and thus is often restricted to a few candidates. Since these candidates are selected merely due to their statistical significance without knowing their identity, their biological interpretation is postponed until further follow-up targeted experiments for structural validation have been conducted. Our computational method integrates potential metabolite ID in the early stages of the workflow and thus allows us to exploit biological information before designing further experiments. Although our method can yield multiple hypothetical IDs for each measured signal, we showed that by combining orthogonal criteria (rt filtering, enrichment analysis) the number of false positive IDs can be cut down significantly and the suggested pathways are reasonable within the scope of discovery studies. For instance, rt prediction models take the physicochemical properties of candidate metabolites into account that are otherwise neglected in an AMS search. By this means, the number of false positive IDs can be decreased tremendously.

In this work, we focused on the analysis of untargeted metabolomics data exclusively and developed a discovery workflow. We plan to extend our workflow by targeted analysis to validate the proposed results automatically.

CHAPTER 6

Untargeted Metabolomics Analysis for Metabolic Diseases

This chapter is based on the same experimental design as previously published in Scientific Reports [19].

6.1 Introduction

Type 2 diabetes mellitus (T2DM) is a metabolic disorder that may cause a multitude of complications such as cardiovascular disease, nephropathic kidney failure, blindness, and the diabetic foot syndrome. These complications are the consequence of *chronic hyperglycemia*, i.e., permanently elevated levels of blood sugar. Although the symptoms of T2DM are well-documented since the ancient times, the importance of *insulin* in glucose homeostasis was first discovered in the 20th century. Insulin belongs to the class of peptide hormones and is secreted by the pancreatic β cells to the bloodstream during the digestion of food carbohydrates. Upon binding to the insulin receptor of most body cells (in particular, adipocytes, skeletal muscle, and liver cells), it initiates the uptake of glucose from the bloodstream through membrane-bound transporter proteins. Simultaneously, metabolic processes to utilize glucose such as glycolysis, glycogen and fatty acid biosynthesis are triggered within the cell. In T2DM, either the efficacy of insulin is decreased (e.g., insulin resistance of body cells) or the insulin production of the pancreatic β cells is impaired, thus causing hyperglycemia. Although this indication was formerly linked to adult humans (*adult-onset diabetes*), it is now known that the classification by age alone is inaccurate since the disease is observed frequently in young humans as well. Instead, new findings suggest that T2DM is strongly correlated with obesity and physical inactivity in consequence to the grave lifestyle changes throughout the last century [170].

In recent years, new evidence was found that T2DM is affected by an individual's genetic background as well. A single nucleotide polymorphism (SNP) located in the TCF7L2 gene (dbSNP entry rs7903146 [171]) was shown to be strongly correlated with the disease [172]. Its phenotype is involved in reduced insulin production [173] and impaired insulin secretion [174, 175]. Functional studies revealed the importance of TCF7L2 to the proliferation of β cells [176, 177]. Beyond its function in the pancreatic islets, TCF7L2 plays regulatory roles in a multitude of glucose-metabolizing tissues such as the liver [178], adipoetic tissue [179], the brain, and the intestine [180]. However, its precise mechanism of function and its influence on the human metabolome in the light of the T2DM pathophysiology remains difficult to explain [181].

Metabolomics studies of SNP-associated disease phenotypes hold great potential to reveal unknown factors and to elucidate the underlying pathophysiological mechanisms. This new knowledge could then be turned into personalized therapies (e.g., lifestyle interventions) and novel diagnostics that would confirm the risk of disease both on the genetic and metabolomic level. In recent years, there has been great interest in integrating the information gained by genome-wide association (GWA) studies with metabolomics data [182]. In the first of such integrative investigations, the influence of several SNPs on the lipid metabolism was confirmed and a potential interplay between a SNP in the FADS1 gene and several glycerophospholipids was suggested [183, 184]. The detection of novel associations was facilitated by using the ratios of related metabolites as estimates for enzymatic activity, effectively lowering their p-values. These examples show that the integration of complementary omics data increases the confidence of statistical results.

In this study, we present a complete computational workflow for the detection of novel metabolite biomarkers predictive of metabolic diseases. Aside from methods for metabolite quantification (Chapter 3) and identification (Chapter 4), our workflow employs a robust statistical approach that is based on RankProd [18] and facilitates the sensitive detection of subtle metabolite fold changes. With our computational workflow, we investigated the impact of the TCF7L2 rs7903146 polymorphism on the human plasma metabolome. To this end, we conducted both clinical chemical and untargeted metabolomics experiments based on a cohort of 30 individuals (15 homozygous non-risk and 15 risk allele carriers). We found several promising candidate biomarkers that suggest perturbations in the steroid and bile acids metabolisms, two well-known implications of T2DM.

6.2 Methods

6.2.1 Study Design

The study design was adapted from our work previously published in Scientific Reports [19].

A cohort of 30 individuals was composed from a database that was previously established in the context of the Tuebingen family (TUEF) study [185]. Our selection excluded all cases with an elevated CRP level ($> 0.5\,\mathrm{mg\,L^{-1}}$), medication use, and elevated levels of type 1 diabetes mellitus (T1DM)-related antibodies (anti-GAD2). Furthermore, only subjects showing an isolated impaired glucose tolerance (iIGT) were picked (fasting plasma glucose $\leq 5.6\,\mathrm{mmol\,L^{-1}}$, 120 min post-challenge glucose between 7.8 and $11.1\,\mathrm{mmol\,L^{-1}}$), resulting in 183 candidates. While ensuring balanced gender numbers and matching mean age and body mass index (BMI), we drew 15 homozygous non-risk (CC) and 15 risk allele (TT) carriers (Table A.10). Our study protocol was approved by the institutional review board of the University of Tuebingen as compliant to the Declaration of Helsinki. All subjects gave their written informed consent to the study. Our investigations followed the ethical principles of good clinical practice.

6.2.2 Bioinformatics Analysis

The raw data files as recorded by the mass spectrometry (MS) software were first converted to the open mzML data standard [57] via the ProteoWizard software (version 2.2.3036) [58]. Both sets of positive and negative mzML files were processed separately with our quantification pipeline (Figure 2.3); the individual processing steps were configured as given by the respective parameter files (Appendix A.1.4). The resulting consensus matrix representations were stored as tabular text files.

All following statistical analysis steps were performed within the R environment (version 2.15.2) [8]. For each row/feature in the consensus matrix, we set a unique row label constructed by the concatenation of the consensus feature coordinates retention time (rt), mass-to-charge ratio (m/z), and intensity (e.g, 54.9_181.17293_1020312). Matrix columns corresponding to the individual feature intensities were extracted, normalized, and filtered as described in Section 2.3.1. Based on the matrix layout as presented in Section 2.3.1, we implemented several filter functions to further condense the dataset. First, we discarded all features that did not show at least twelve non-missing values in one of the subject groups (80 % rule) [15]. From this prefiltered set, we further removed rows if the median level of their individual intensities were below threefold of the corresponding blank intensities. To assess the reproducibility of replicate feature intensities, we computed the coefficient of variation (CV) for each row based on the quality control (QC) measurements. Features showing a CV higher than 30 % were removed from the dataset [15]. Finally, we stripped the intensity columns corresponding to the blanks and QC measurements, gaining the data matrix ready for statistical analysis.

This dataset was logarithmized to the base of two and processed with the RankProd R package [102] (Section 2.3.2). Class information was supplied via a vector of class labels (0 and 1 designating the non-risk and risk alleles, respectively). The number of permutations was set to 5,000. From the resulting data object, we extracted all significant features with a percentage of false positives (PFP) (used synonymously to false discovery rate (FDR)) up to 20 % and stored these together with additional information (e.g., p-values and fold changes) in a tabular text file. Since in some cases the RankProd procedure did not report the fold change, we additionally computed the ratio between the median levels of the risk and non-risk groups. Based on the list of the significant features' rt and m/z coordinates, our mapCutter.sh command-line script was employed to cut out the corresponding extracted ion chromatogram (XIC) from the original mzML input files (see appendix A.1.4). The excised chromatographic profiles were superimposed in R plots with respect to their class in order to compare them visually against the blank profiles. For convenience, a group-wise box plot was generated to visualize the group differences in a quickly accessible manner. We automated the generation of these four-panel plots with our R script plotGroupwiseXICs.R (Appendix A.1.4).

We performed an AccurateMassSearch (AMS) run with relative isotopic abundance (RIA) filtering to putatively identify the significant features (Section 4.2.2). Both for the positive and negative mode data, the identification was restricted to the most common and simple adducts (for parameter settings, see Appendix A.1.4). Putative metabolites corresponding to exogenous sources (e.g., food or drugs) were filtered out. Each feature was checked visually in the original data (e.g., if a clear elution profile was observable) to rule out spuriously detected features stemming from contaminant compounds. To further confirm our putative IDs, we studied the relationship between the compounds' observed rts and their predicted hydrophobicity (XLogP) values in a robust regression analysis (Section 4.2.5). The positive and negative mode IDs were considered in separate regression analyses. Duplicate data points (i.e., identical rt and XLogP values) were removed from the training datasets. Metabolite IDs were marked as unlikely if they were outside the middle 95 % interval of the residual distribution.

6.2.3 Untargeted Metabolomics Analysis

The following sample preparation and ultra performance liquid chromatography (UPLC)-MS analysis steps were performed by Jia Li (CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, China) and were previously described in [19].

In order to prepare the ethylenediaminetetraacetic acid (EDTA) plasma for the LC-MS measurements, a 200 µL aliquot was deproteinized with acetonitrile (target concentration 66 %), vortexed for 2 min, and centrifuged for 20 min at 15,700 g and 4 °C. The supernatant was then dried in a vacuum centrifuge and stored at −80 °C. Immediately before the analysis, the samples were reconstituted in a 200 µL acetonitrile/water mix (*v*/*v* 4:1).

The metabolomics profiling experiments were performed on an Accela^{TM} UPLC system coupled to an LTQ-Orbitrap XL mass spectrometer with an electrospray ionization (ESI) source (Thermo Fisher, San Jose, CA, USA). For the chromatographic separation, we employed a 2.1 mm×100 mm×1.7 µm C_8 AQUITY^{TM} column (Waters, Milford, MA, USA). The separation protocol was described previously [186] and was changed slightly for our experiments. In the positive mode analysis, we used water with 1 % formic acid (A) and acetonitrile (B) as mobile phases. The elution gradient was initialized with 95 % A and 5 % B for 0.5 min, increased linearly to 100 % B for the next 23.5 min, maintained for 4 min, and finally returned to 5 % B within 0.1 min. Before the next sample injection, the column was equilibrated for 4 min.

The chromatography runs were carried out with a flow rate of 0.35 ml min and a column temperature of 35 °C. In negative mode analysis, the mobile phases were water (A) and 95 % MeOH/5 % H_2O (B), both with 5 mmol NH_4HCO_3. The elution gradient began with 20 % B for 0.5 min, increased linearly to 100 % B for the next 21.5 min, maintained for 5 min, and finally returned to 20 % B within 0.1 min. Column equilibration was performed for 5 min before the next injection. Flow rate was set to 0.35 ml min and the column was operated at 50 °C.

The LC capillary temperature was at 325 °C. In positive mode, the ESI source was operated with a voltage of 4.5 kV, sheath gas flow at 40 arb. unit and auxiliary gas flow at 5 arb. unit. In negative mode, the source voltage was set to 3.5 kV, the sheath gas flow to 35 arb. unit and auxiliary gas flow to 5 arb. unit. High resolution full scans were recorded between 100–1,000 Th and data was acquired in profile mode. Samples were measured in a randomized order. Every fifth injection, a QC sample was analyzed to assess reproducibility [187]. After every tenth injection, a blank sample was measured.

6.2.4 Clinical analyses

The following analyses were previously described in [19].

Following a 12 h overnight fasting, we performed a 75 g oral glucose tolerance test (OGTT) with each subject at 8 am to investigate the kinetics of glucose, insulin, C-peptide, proinsulin, and non-esterified fatty acids (NEFA). Blood glucose levels were measured with a bedside glucose analyzer (YSI, Yellow Springs, CO). Insulin and C-peptide were determined on an ADVIA Centaur XP, blood cell count on an ADVIA 2120, and the remaining routine parameters on an ADVIA 1800 clinical chemistry system (all devices from Siemens Healthcare Systems, Erlangen, Germany). The genotyping of the rs7903146 SNP was performed on the MassARRAY platform (Sequenom, San Diego, USA). We calculated the insulin sensitivity from the glucose and insulin levels with the method proposed by Matsuda and DeFronzo [188]. Insulin secretion was estimated from glucose and C-peptide levels according to the homeostasis assessment model (HOMA) [189] via a freely available calculator software [190].

6.3 Results

The statistical analysis of the positive mode data reported 60 features that showed significant differences in their levels between the control and risk allele groups (35 up- and 25 down-regulated, see Tables A.11 and A.13, respectively). Among these, we could assign a unique empirical formula with at least one putative Human Metabolome Database (HMDB) metabolite ID to five up-regulated and eight down-regulated candidates (Tables A.12 and A.14). The list of putative IDs comprised lipid compounds (phospholipids, FFAs), bile acid glycine conjugates, sugars, and the amino acid *L-tyrosine*. Most of these IDs were quite reasonable in the context of T2DM and were reported in the literature (e.g., FFAs and phospholipids [191] or bile acids [20]).

One interesting up-regulated candidate (m/z 181.07109) could be mapped to several isobaric hexose sugars with *glucose* being the most prominent one. While the fold change between this candidate's group-wise median levels was clearly observable (see box plot in Figure A.18), the classical approach of comparing both groups with a two-sample t-test yielded no significant difference. We found that RankProd was still sensitive enough to detect the differences between the groups and to rank this as a potential candidate biomarker. In general, the fold changes reported by the RankProd function were rather small as expected for metabolites (1.20 to 1.53 for up-regulated and 0.67 to 0.86 for down-regulated candidates), and with rather high biological variances present in the groups, common two-sample t-tests were not robust enough to detect these subtle differences. As a counterexample, the group differences of the down-regulated candidate putatively identified as *N-undecanoylglycine* (Table A.14 (3)) were much clearer both in the corresponding group-wise XIC plots and intensity box plots (Figure A.19) and also significant according to a t-test (p-value $= 0.021$, $\alpha = 0.05$). The function of N-undecanoylglycine, an odd-numbered acylglycine, is unknown in the T2DM context but might indicate an impairment of the lipid metabolism and thus is a promising biomarker candidate.

The negative mode dataset yielded 48 candidates with significant fold changes (22 up- and 26 down-regulated, see Tables A.15 and A.17). We could assign a unique empirical formula with at least one putative HMDB metabolite to ten up- and nine down-regulated candidates (Tables A.16 and A.18). In particular, these putative IDs comprised several compounds from the lipid class (di- and triglycerides, phospholipids, FFAs), bile acids (deoxycholic acid (DCA) and chenodeoxycholic acid (CDCA)) and their conjugates, and purines. For instance, two candidates were putatively identified as *pentacosanoic* and *hexacosanoic acid*, two FFAs (Table A.16 (4) and (10)). FFAs are interesting markers for the lipid metabolism and were studied frequently in the context of T2DM [192, 193, 194]. Again, we found N-undecanoylglycine as a down-regulated candidate with a similar fold change of 0.69. More strikingly, one candidate was identified as *androsterone* or *dihydrotestosterone (DHT) sulfate*, two isobaric sulfated breakdown metabolites of the androgenic steroid hormone *testosterone* (Table A.18 (4) and Figure A.20). In the same context, another very interesting finding was the candidate putatively identified as *androstanediol glucuronide* (Table A.18 (9) and Figure A.21). Both candidates were detected as down-regulated in the risk allele group and might be interesting biomarkers for the impairment of testosterone metabolism, a common implication of T2DM. In the group-wise XIC plots and intensity box plots, the difference in the group median levels was observable, however, the risk allele groups were not significantly lower in their levels according to a one-sided Wilcoxon rank sum test (α=0.05). To further assess the likelihood of these IDs, we broke the cohort and the corresponding measurements down not only by genotype but also by gender. As expected, the levels of both markers were significantly lower in the female subjects than in the male ones (Figure 6.1). Furthermore, the difference between the non-risk and risk allele groups were now more evident and in some cases also statistically significant. This was particularly true for the androstanediol glucuronide marker. In case of androsterone/DHT sulfate, however, we observed high biological variances within the risk allele groups such that the difference between

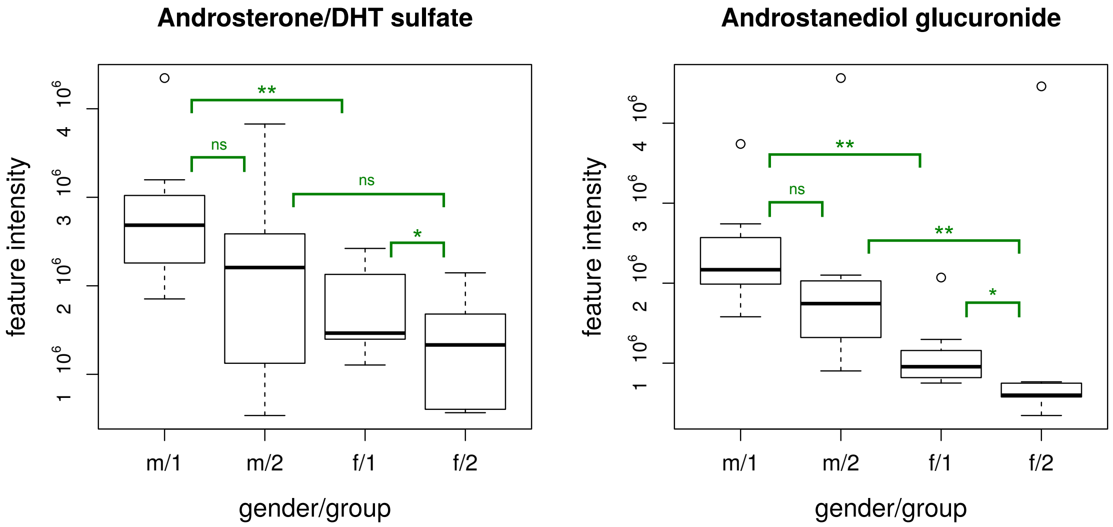


Figure 6.1: Levels of androsterone/DHT sulfate and androstanediol glucuronide with respect to gender/group. The box plots show the feature intensity distributions of the individuals grouped by gender and genotype (group 1 for non-risk, group 2 for risk allele group). We performed pair-wise one-sided Wilcoxon rank sum tests to assess if the conditions compared are significantly different in their levels or not (see green brackets). ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$.

groups could not be confirmed as statistically significant. We double-checked if a similar dependency on gender existed for the other candidates (e.g., N-undecanoylglycine) but found no evidence as strong as in these two cases. Based on our observations, we considered these two IDs as highly reliable markers that should be targeted first in follow-up experiments.

With regard to the rt prediction models based on the putative IDs from positive and negative mode, respectively, we found robust linear models that could describe the observed rts very well (Figure 6.2). The rt model trained with the positive IDs data achieved an R^2 of 0.948 and an RMSE of 2.33 min (values averaged over 500 runs of tenfold cross validation). The underlying residual distribution revealed four outlying data points. These corresponded to 19 potentially false positive identifications of which 16 belonged to phosphatidylcholines (PCs) isomers (Table A.14). In case of the negative mode ID, the rt model achieved a cross-validated R^2 of 0.868 and an RMSE of 2.88 min. Four data points were marked as outliers corresponding to ten potentially false positive IDs (Tables A.16 and A.18).

6.4 Discussion

The statistical analysis of the LC-MS measurements yielded 108 potential markers suggesting that there are differences in metabolite patterns between the non-risk and risk allele groups. With our metabolite identification workflow, we could annotate 32 of these candidates with a unique empirical formula and corresponding putative IDs from the HMDB database. Although several marker masses could not be identified unambigously (i.e., not matched uniquely with

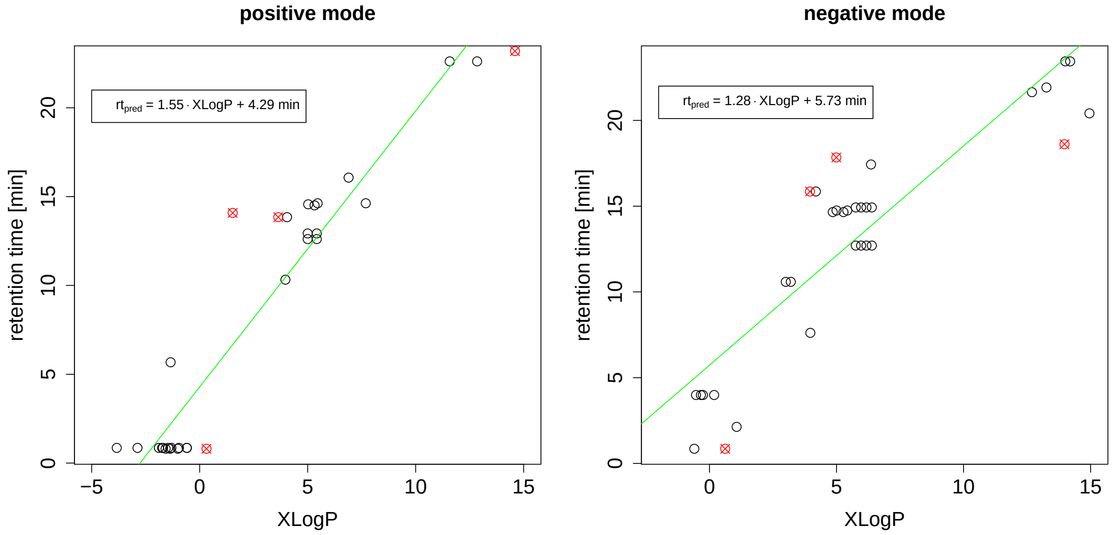


Figure 6.2: Retention time models based on the TCF7L2 dataset. Robust regression models were trained to explain the rts of metabolite marker candidates from the positive and negative modes, respectively. Both models exhibit four outlier data points that most likely stem from false positive putative IDs (Tables A.14, A.16 and A.18).

one metabolite), the isobaric alternative identifications were usually from the same compound class and thus suggested the potential role of specific metabolic pathways. For instance, in the negative mode dataset we found metabolite ions that were exclusively annotated with isobaric bile acids or their conjugated derivatives (Table A.18). Other ambiguities were introduced by isobaric lipids with varying double bonds, steroid hormons and their conjugated forms, and hexose sugars or their breakdown products (e.g., glucose or lactic acid). By harnessing the orthogonal information of the candidates' observed retention times with rt prediction models, we could rule out 29 putative HMDB IDs. In some cases, this left metabolite ions with no ID at all. For example, heptanoylcarnitine (Table A.14 (1)) appeared less probable according to the rt model. Since the metabolite ion had a quite high rt (14.13 min) and the postulated adduct formation with potassium appeared quite unlikely, marking this ID as invalid was sound. In other cases, metabolite ions could be uniquely annotated after the alternative IDs were rendered invalid, e.g., trans-dec-2-enoic acid (Table A.14 (5) or 3,17-Androstanediol glucuronide (Table A.18 (9)). In general, the rt models showed that most putative IDs suggested by our AMS and RIA workflows correlated excellently with the observed metabolite features and thus helped to improve the confidence in our computational identification results.

Under the assumption that most metabolites were correctly identified, we studied some marker candidates and their potential role in T2DM. Several IDs suggested the involvement of the bile acid CDCA together with its glycine and taurine conjugates. In negative mode, these markers were found as down-regulated with similar fold changes around 0.82, i.e., the decrease of CDCA would affect its conjugates to the same extent. The relationship between these markers might suggest that both the production of CDCA, one of the primary bile acids produced in the human body, and its downstream conjugation might be impaired in the risk-allele group. Another supporting evidence could be an up-regulated marker with a fold change of 1.33 (negative mode) that was putatively identified as 3a,7a-Dihydroxycholanoic acid, a precursor bile acid upstream of the CDCA synthesis. This might indicate decreased enzyme activities along the conversion path from cholesterol to CDCA. Since the same pool of precursors leading to CDCA also feeds the synthesis of the other primary bile acid cholic acid (CA), we expected to see perturbations in the levels of CA or its downstream bile acids as well. An altered CA/CDCA ratio is believed to be associated with the T2DM disease state [20] and thus would make sense in the context of the TCF7L2 rs7903146 polymorphism. Another common complications of T2DM are sexual dysfunctions, usually associated with impaired levels of testosterone [22, 21]. Our negative mode data indicated the presence of two testosterone metabolites, namely androsterone or DHT sulfate and 3,17-Androstanediol glucuronide, that were down-regulated. 3,17-Androstanediol is the major metabolite of DHT and both are known to play crucial roles in sex drive and function [195]. When we separated the groups additionally by gender, the expected differences between male and female androgen metabolite levels became more clear and thus provided a strong evidence for the validity of these IDs. Finally, we found potential FFAs markers which were increased by a factor of roughly 1.5 in the risk-allele group (Tables A.15 and A.16 (4) and (10)). Elevated FFAs plasma levels were linked to insulin resistance and impaired β cell function [192, 193, 194]. With respect to the β cell function, we could confirm this by a HOMA index lower in the risk-allele group than in the control group (Table A.10). However, no significant differences regarding insulin resistance were measured in our cohort.

This study on the TCF7L2 rs7903146 polymorphism revealed several potential biomarkers that discriminated well between the non-risk and risk allele groups. Interestingly, these differences were often subtle since the fold changes were rather small (1.15 to 1.53 for up-regulated and 0.54 to 0.86 for down-regulated candidates). This was complicated by the fact that many of the candidates showed a high biological variation within each group and thus the group-wise intensity distributions were strongly overlapping. In our earlier work based on the same dataset [19], no significant perturbations on the metabolome level could be detected. The feature dataset was extracted with the vendor MS software SIEVE [196] and the analysis was performed with PCA and PLS-DA, statistical tools that are very common in the metabolomics community. However, our present work showed the RankProd method combined with our signal processing pipeline was more adequate to detect such subtle metabolite perturbations than the classical data processing and statistical analysis tools. Since the outcomes of our former and latter approaches were quite different and allowed for contradicting interpretations, this comparison clearly demonstrated the impact of data processing and statistical techniques on the fate of biomarker studies. A preliminary interpretation of our data was facilitated by our automated metabolite annotation workflow yielding confident IDs that could be explained plausibly in the given biological context. Nonetheless, these IDs must be confirmed by a targeted approach using standard compounds or MS/MS fragmentation before final conclusions about the pathophysiological implications can be drawn. To this end, our list of putative IDs could guide the design of targeted follow-up experiments.

We plan to conduct targeted experiments in order to elucidate the structure of the most promising putative IDs. Particularly, a deeper investigation of the metabolites involved in the bile acid, steroid, and lipid metabolism seem to be most promising. Furthermore, we would like to explore alternative parameter sets in the metabolite IDs workflows (e.g., more complex adduct formations) since this could yield more marker candidates that were missed with our present conservative settings. Finally, the integration of further orthogonal criteria (e.g., simple chemical transformation rules such as water loss or decarboxylation in order to link related adducts) might lead to more confident identifications.

CHAPTER 7

Conclusion

As we have shown in this work, we made valuable contributions to the field of computational metabolomics. We designed and implemented novel algorithms and data analysis workflows to tackle its central problems: the automated quantification of metabolites, their identification, statistical analysis, and ultimately their biological interpretation. Our main intention was to solve these problems in a comprehensive and consistent manner rather than addressing them separately. This high level of consistency sets our computational workflows apart from existing solutions that usually address subproblems only and are difficult to integrate with each other. Still, our workflows offer high flexibility due to their modular design in contrast to monolithic all-in-one solutions. Each of our algorithms was implemented as a self-contained *module* that can be easily swapped or reused in another workflows. This flexibility allows us to quickly adapt to the rapidly evolving requirements of high-throughput metabolomics experiments.

Our computational strategies facilitate the rapid extraction of novel biological knowledge from complex LC-MS data. We could prove this by producing biologically sound results on the basis of the TCF7L2 polymorphism dataset (Chapter 6). We developed a comprehensive data analysis workflow for biomarker discovery in untargeted LC-MS data that built upon the statistically robust *rank product* test [18] and included a thorough data pretreatment strategy geared towards metabolomics experiments (Section 2.3.1). In contrast to our previous work [19], we found more than 100 potential biomarkers that suggested significant differences between control and risk allele subjects and thus a strong involvement of the studied TCF7L2 polymorphism in T2DM. Some biomarker candidates indicated perturbations in the bile acid and androgenic steroid biosyntheses, both of which were studied extensively in the same context before [20, 21, 22]. We concluded from our previous and present results that the choice of data preprocessing and statistical methods were critical to detect the subtle changes in metabolite levels between control and treatment groups. This was particularly true for metabolites that exhibited strong intra-group variation [15]. In our previous work [19] we employed PLS-DA, a popular statistical tool in the metabolomics community to detect discriminant metabolite patterns. However, it has been shown that it might not be sensitive enough to capture small differences in metabolite levels without proper optimization [17]. In this regard, we found that our biomarker discover workflow was more robust and sensitive in the detection of novel metabolite biomarkers than traditional methods such as PLS-DA.

Temporally resolved metabolomics data is another application field where classical approaches such as PCA and PLS-DA are not readily applicable [148]. To this end, we implemented a novel computational pipeline that boils down complex time-course metabolomics data to a few characteristic metabolic patterns and correlates them with pathway information to assist in biological interpretation (Chapter 5). The centerpiece of this pipeline is a variant of the QT clustering algorithm [23] which we extended with the STS distance [24]. In contrast to the standard Euclidean distance or Pearson correlation as employed in most clustering approaches, the STS metric offers higher discriminatory power when comparing time series with non-equidistant time points. Furthermore, it is particularly suited to time series with only few time points [24]. This is a very common scenario in metabolomics studies where sample material is often limited (e.g., human samples, biopsies). We found that our modified clustering algorithm captured the rapid changes during short time intervals more precisely while allowing for a bigger biological variation in longer intervals. We validated our clustering approach on the basis of time-course data that was acquired during a single bout of exercise. Mapping the two most prominent clusters to the SMPDB revealed several pathways related to amino acid, FFA, and catecholamine metabolism. In literature, these pathways often occur in the context of physical exercise [162, 163, 168]. These results clearly showed that our discovery pipeline is a potent tool to elucidate characteristic kinetic patterns from untargeted time-series data. In a follow-up experiment, these patterns can be further studied in a targeted approach, saving both time and cost.

With our biomarker discovery pipelines, we achieved very promising results that allowed for a sound biological interpretation of LC-MS data. However, this would have not been possible without proper metabolite ID. We developed an integrative strategy for metabolite ID that exploits all sources of orthogonal information extractable from untargeted LC-MS data (Chapter 4). Given the m/z of an observed metabolite feature, our AMS tool can detect a wide range of potential positive and negative adducts. We observed that the more potential adducts were considered during an AMS run, the more the actual metabolite IDs were overshadowed by false positives. In order to reduce the ambiguity and to increase the confidence of our ID results, we implemented several orthogonal filter strategies. RIA filtering can significantly lower the number of empirical formulas that were suggested by spuriously detected adducts. The effectiveness of RIA filtering was shown before [65, 66] and could be confirmed by our experiments (Section 4.3.1). Obviously, the presence of a metabolite's isotope pattern is required by the RIA filter — a condition that is rarely met for low-concentrated metabolites close to the LOD. Another filter criterion employs an rt prediction model to assess if proposed candidate IDs are close to the observed feature's rt. We devised a computational workflow to train such rt models automatically on the basis of confirmed metabolites (Section 4.2.5). Our experiments suggested that rt filters are very effective in discarding false positives while increasing the confidence in the correct metabolite IDs (Section 4.3.4).

Ultimately, following the guidelines of the MSI [12], the strongest evidence of a metabolite ID is provided by a match against a spectral database of authentic standards measured under identical conditions. We designed the MSM tool to facilitate efficient matches against the public fragment database MassBank. While MSM can easily be extended by fragment data other than MassBank, the availability of public resources for offline use is very low. Common small molecule fragment databases such as the METLIN database solely provide search functionality via a web-based interface [85]. Although the coverage of metabolites found in MassBank is incomplete, we observed excellent results in a RPLC-MS/MS validation dataset. In almost all cases, MSM ranked the correct metabolite IDs as best hits (Section 4.3.2). Furthermore, the matching scores were significant according to the background score distribution. We adapted MSM's scoring function from the dot product based HyperScore [82] which was employed in the X!Tandem search engine for peptides [83]. With respect to fragment data from small molecules, similarity measures with better discriminatory power than the dot product have been proposed [84]. While MSM currently provides only one scoring function, it can easily be augmented by alternative metrics.

In order to harness the advantages of both AMS with orthogonal filters and MSM, we merged the two tools into a combined ID approach. According to our validation experiments (Section 4.3.3), AMS tends to detect more authentic standards than MSM but for the price of less specificity. On the contrary, MSM discriminates well between the correct and false positive IDs but strongly depends on the metabolite coverage of its underlying spectral DB and whether a precursor ion was selected by DDA or not. Based on our visual annotation of the validation dataset, we found that metabolite IDs confirmed or excluded by both AMS and MSM (44 out of 54) were highly reliable. Candidates that were proposed only by AMS could be trusted if they were confirmed by one or more orthogonal criteria. However, none of the orthogonal filters applied in these cases. Still, from our assessment of the individual RIA and rt filter performances we would expect an improvement of the ID accuracy if these were applicable (Sections 4.3.1 and 4.3.4).

In an untargeted metabolomics experiment, the very first step is the quantification of all metabolites present in a complex biological sample. Hence, this step is critical for the quality of following analyses and ultimately for the final results. We developed the robust feature detection algorithm FFM that is sensitive to low-intensity metabolites close to the detection level and offers an excellent correlation between spiked concentrations and observed intensities of authentic standards in human plasma (Chapter 3). Aside from a sensitive detection of mass traces, we introduced a novel SVM classifier for deisotoping. Our classifier was built on the basis of isotope patterns that were generated with computational means [66] to capture the domain of small organic molecules exhaustively. We evaluated its performance with data from an actual metabolite DB and were able to differentiate between real metabolites and invalid isotope patterns with outstanding performance. In a benchmark with simulated LC-MS datasets, we compared the performance of FFM with one of the state-of-art algorithms XCMS/CAMERA. Our algorithm clearly exceeded the XCMS/CAMERA performance in terms of recall and precision, mainly due to a much higher sensitivity to low-intensity features. Implemented as a self-contained module within the OpenMS framework, it is easy to integrate into more complex quantification (signal preprocessing, feature linking), identification, and statistical analysis pipelines.

In future work, we would like to take the promising results obtained with our discovery pipelines one step further and validate them in targeted metabolomics experiments. With respect to the TCF7L2 dataset, the confirmation of the biomarker candidates pointing at the bile acid, steroid, and lipid metabolism could shed light on the controversy whether the rs7903146 polymorphism has a real impact on T2DM or not [19, 176]. In the same line, we plan to experimentally validate the metabolic pathways that were revealed by our enrichment analysis and most likely were perturbed during the exercise and recovery stages of our time-course experiment. Metabolite ID was performed with very conservative settings in order to keep the number of false positives low. By doing so, however, we might have missed some potential biomarker IDs. We would like to study more relaxed parameter settings (e.g., extended range of potential adducts) in conjunction with additional orthogonal filter criteria. Novel filters could integrate more chemical knowledge into the identification process, e.g., common chemical transformations (dehydration, decarboxylation, deamination). Our AMS could greatly benefit from the integration of adduct likelihoods, that is, chances of specific adducts to be formed during ESI ionization. With respect to metabolite ID by means of spectral matching, neither the TCF7L2 nor the exercise dataset provided MS/MS spectra. To this end, we plan to study RPLC-MS/MS measurement protocols that would allow for both robust quantification and confident ID at the same time. Along that line, we would like to further improve on the MSM performance by increasing the metabolite coverage of our spectral resource. Furthermore, the integration of alternative scoring functions such as the X-Rank score [84] could increase the discriminatory power of spectral matching, in particular, if the spectra are as heterogeneous as in the MassBank case [13] (e.g., different MS platforms and fragmentation energies). Finally, we strive to increase the specificity of our FFM algorithm which could be realized by a more sophisticated signal-to-noise estimation. The performance of our method could be additionally improved by taking the feature reproducibility in multiple measurements (technical and biological replicates) into account. We plan to implement an OpenMS tool to detect and merge adduct features that originate from the same metabolite. By this means, we could further decrease the redundancy present in the feature lists.

APPENDIX A

Supplementary Material

A.1 File Archive

A.1.1 File Listing of Chapter 3

Quantification Dataset

```
./Chapter3
 ├─featlinking.ini
 └─PlasmaData
    └─FFM_QuantBenchmark.ini
```

Simulated Plant Metabolites Dataset

```
./Chapter3
 ├─featlinking.ini
 └─SimData
    ├─detNoiseSim
    ├─errorPPMSim
    ├─GroundTruth.featureXML
    ├─intDistSim
    ├─mssim_template.ini
    └─plant_mets_270212.csv
```

A.1.2 File Listing of Chapter 4

RPLC-MS/MS Validation Dataset

```
./Chapter4
└─primary_data
   └─FFM.ini
```

Accurate Mass Search with Isotope Ratios Filtering

```
./Chapter4
└─AMS
   └─parseHMDB30.py
```

Spectral Search in Fragment Spectra Databases

```
./Chapter4
└─MSM
   └─parseMBFiles.py
```

Exploiting the AMS and MSM Tools in a Combined Approach

```
./Chapter4
└─combined
   ├─mergeIDs.py
   └─inis
      ├─ams_pos_5ppm.ini
      ├─ams_pos_1ppm.ini
      ├─ams_pos_5ppm_onlyH.ini
      ├─ams_pos_1ppm_onlyH.ini
      ├─ams_neg_5ppm.ini
      ├─ams_neg_1ppm.ini
      ├─ams_neg_5ppm_onlyH.ini
      ├─ams_neg_1ppm_onlyH.ini
      └─ams_pos_riafilt.ini
```

Prediction of Metabolite Retention Times on RPLC-MS Columns

```
./Chapter4
└─rtpred
   ├─SimplePositiveAdducts.tsv
   └─SimpleNegativeAdducts.tsv
```

A.1.3 File Listing of Chapter 5

Metabolite Signal Extraction

```
./Chapter5
└─feature_extraction
   └─findFeatures.py
```

Metabolite Identification

```
./Chapter5
└─metabolite_id
   ├─clusterIDmapping.R
   └─inis
      ├─AMS_pos.ini
      ├─AMS_neg.ini
      ├─PositiveAdducts.tsv
      └─NegativeAdducts.tsv
```

Cluster Analysis

```
./Chapter5
└─qt_clustering
   ├─qtClustPipeline.R
   └─flexclust_final.zip
```

Pathway Enrichment Analysis

```
./Chapter5
└─enrichment
   └─hmdbPathwayMapper.py
```

A.1.4 File Listing of Chapter 6

Bioinformatics Analysis

```
./Chapter6
├── ID
│   ├── AMS_neg.ini
│   ├── AMS_pos.ini
│   ├── NegativeAdducts.tsv
│   └── PositiveAdducts.tsv
├── quant
│   ├── FeatureLinkerUnlabeledQT.ini
│   ├── FFM_neg.ini
│   ├── FFM_pos.ini
│   └── PeakPickerHiRes.ini
└── tools
    ├── mapCutter.sh
    └── plotGroupwiseXICs.R
```

A.2 Figures

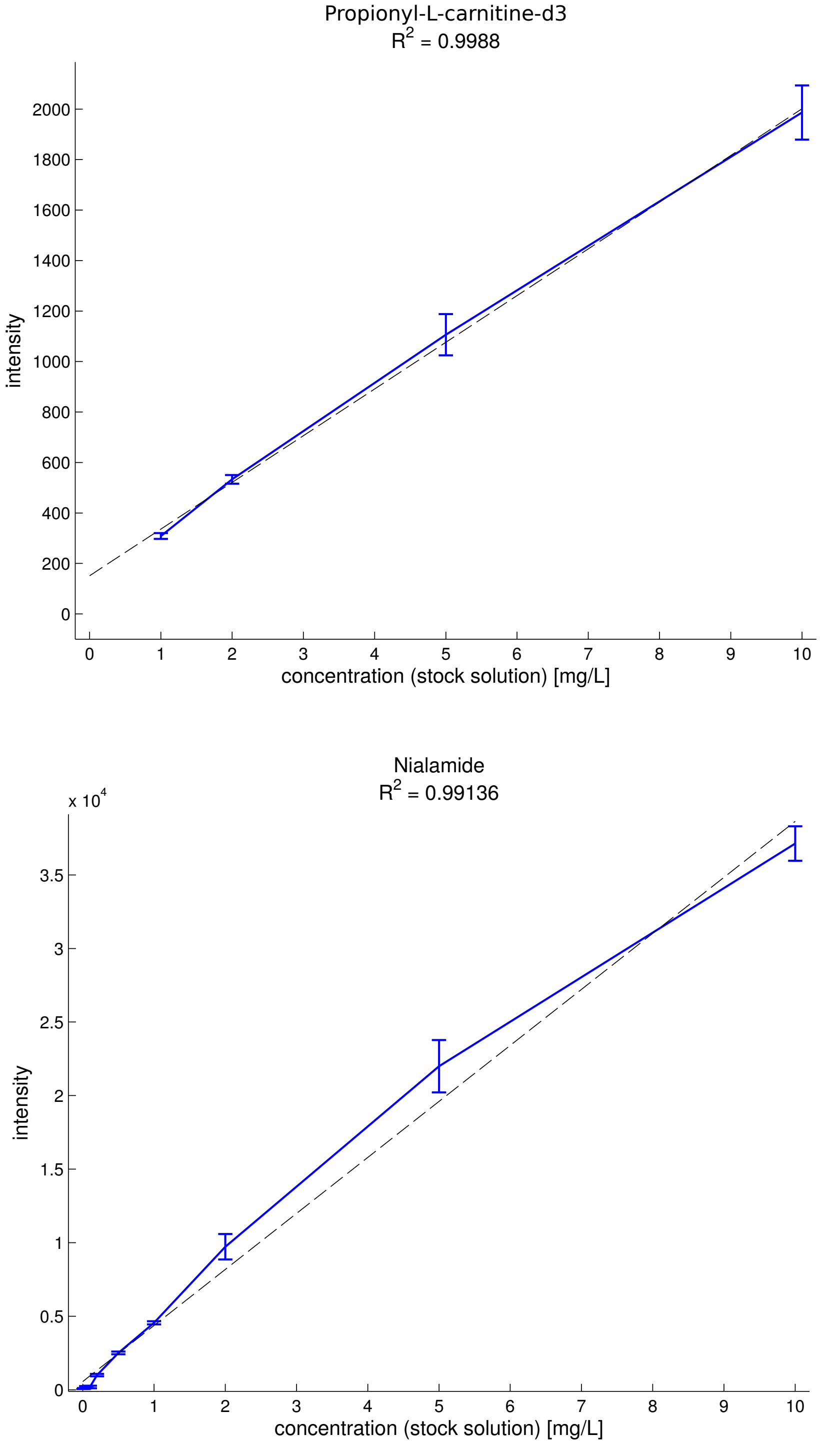


Figure A.1: Relationship between propionylcarnitine-d3 and nialamide concentrations and respective feature intensities. For each concentration, an error bar depicts the standard deviation of the triplicate intensities. The solid line connects the mean intensities obtained from each of the 11 triplicates. For each compound, a line was fit by linear regression to the 33 concentration-intensity data points with the goodness-of-fit R^2 (dashed line) [53].

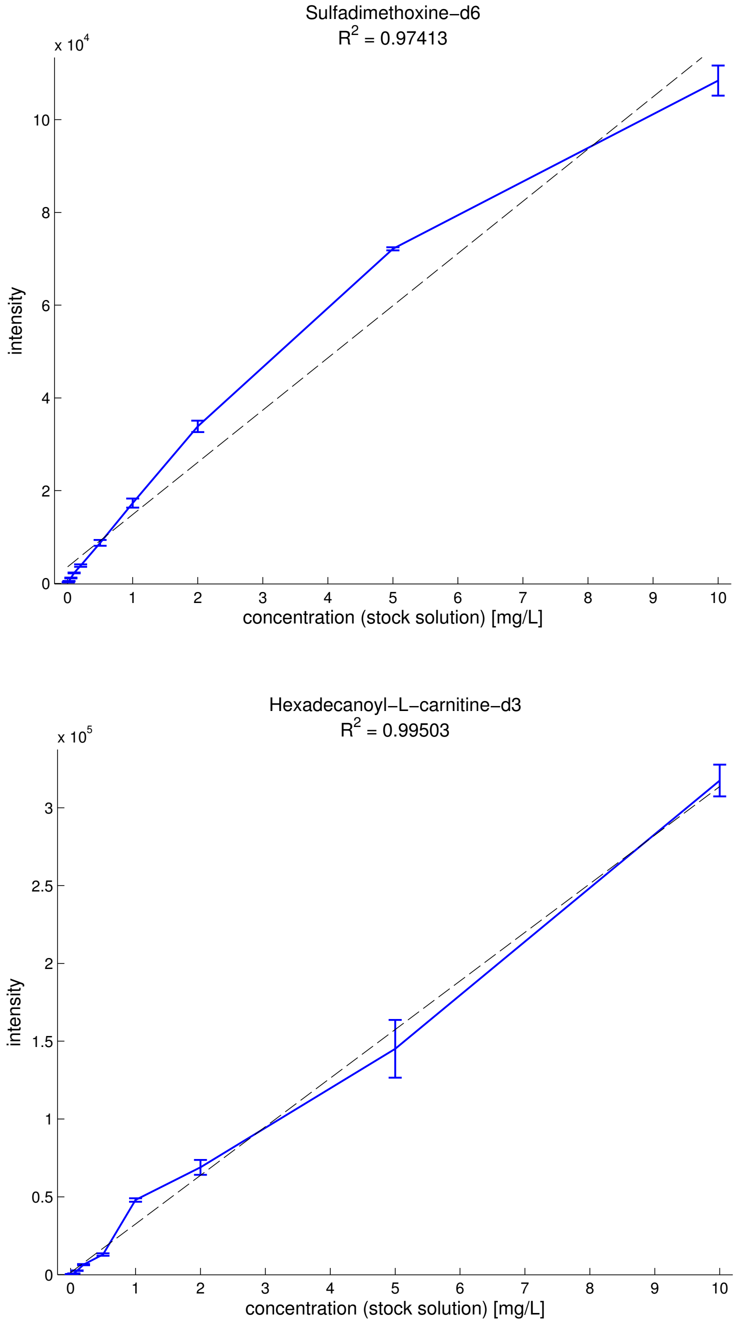


Figure A.2: Relationship between sulfadimethoxine-d6 and hexadecanoyl-L-carnitine-d3 concentrations and respective feature intensities. For each concentration, an error bar depicts the standard deviation of the triplicate intensities. The solid line connects the mean intensities obtained from each of the 11 triplicates. For each compound, a line was fit by linear regression to the 33 concentration-intensity data points with the goodness-of-fit R^2 (dashed line) [53].

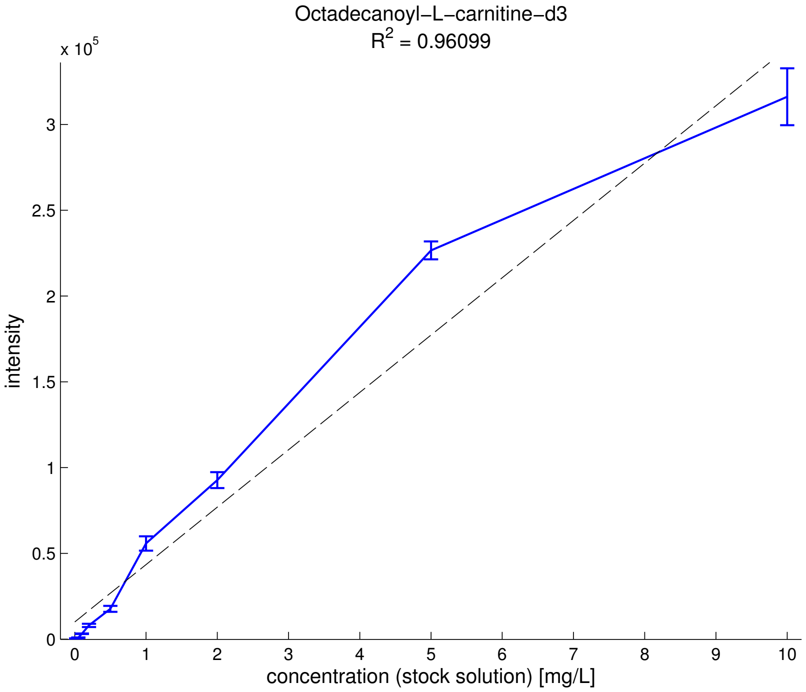


Figure A.3: Relationship between octadecanoyl-L-carnitine-d3 concentrations and feature intensities. For each concentration, an error bar depicts the standard deviation of the triplicate intensities. The solid line connects the mean intensities obtained from each of the 11 triplicates. A line was fit by linear regression to the 33 concentration-intensity data points with the goodness-of-fit R^2 (dashed line) [53].

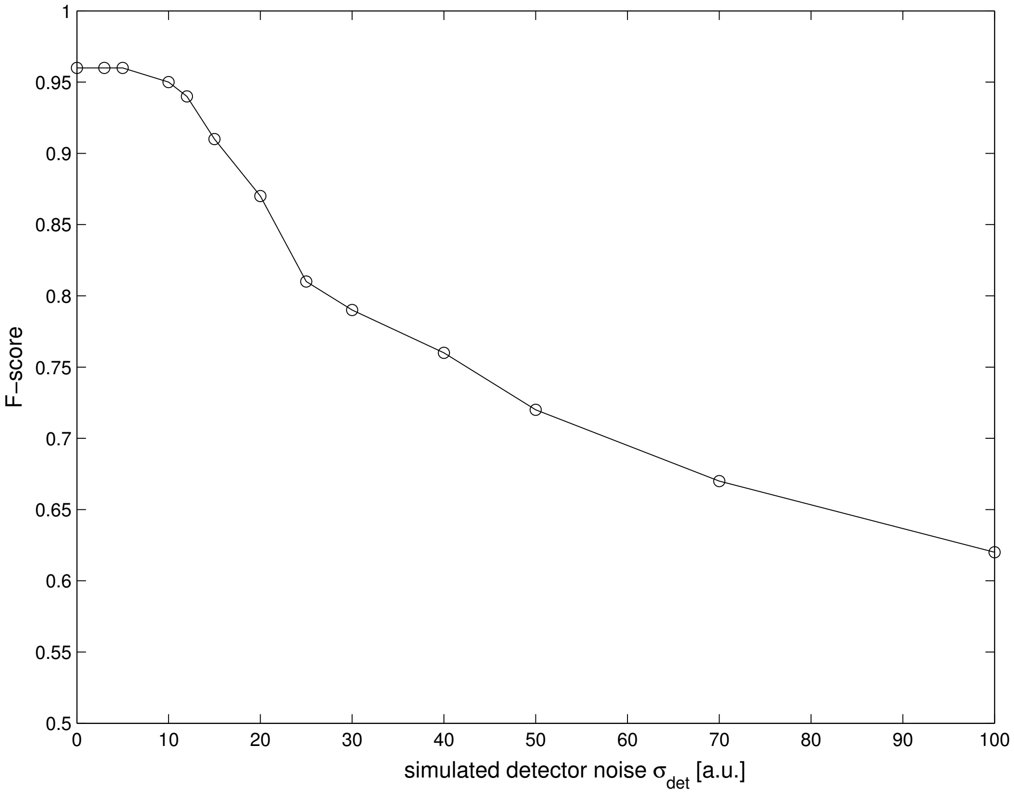


Figure A.4: Feature finding performance with respect to varying degrees of simulated detector noise. For each simulated detector noise setting, the `noise_threshold_int` parameter was chosen accordingly as an optimal setting. Additionally, only mass traces with a signal-to-noise ratio of at least 10 were accepted [53].

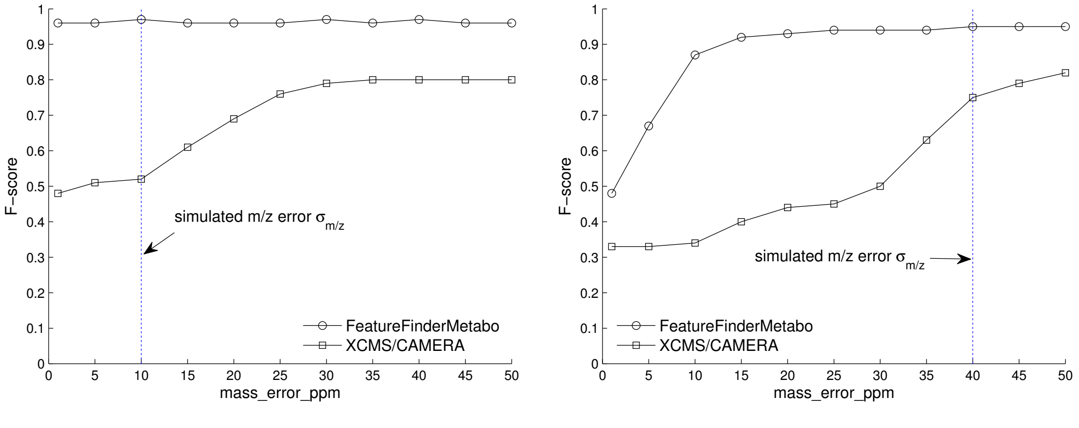


Figure A.5: Feature finding performance comparison with respect to varying settings of the `mass_error_ppm` parameter. Feature detection runs were performed on datasets with a simulated m/z error $\sigma_{m/z} = 10$ ppm (left plot) and $\sigma_{m/z} = 40$ ppm (right plot). Our algorithm's `mass_error_ppm` parameter and the corresponding parameter in XCMS/CAMERA were varied between 1 and 50 ppm, respectively [53].

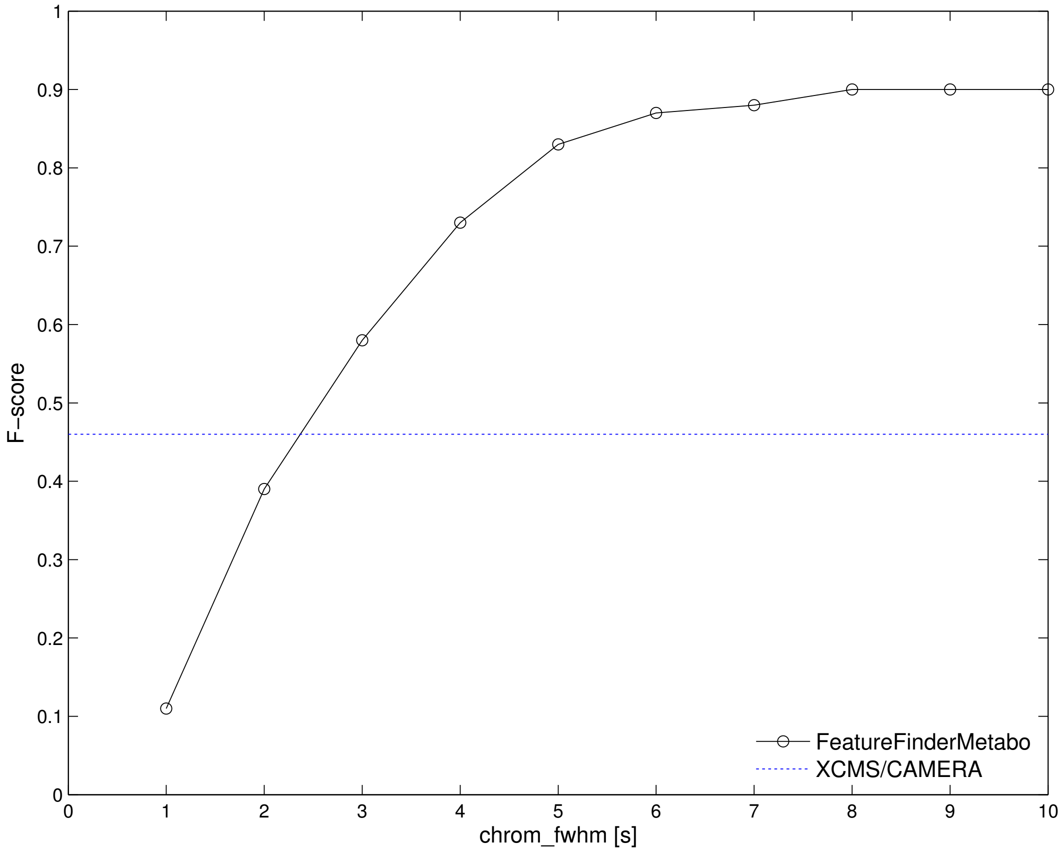
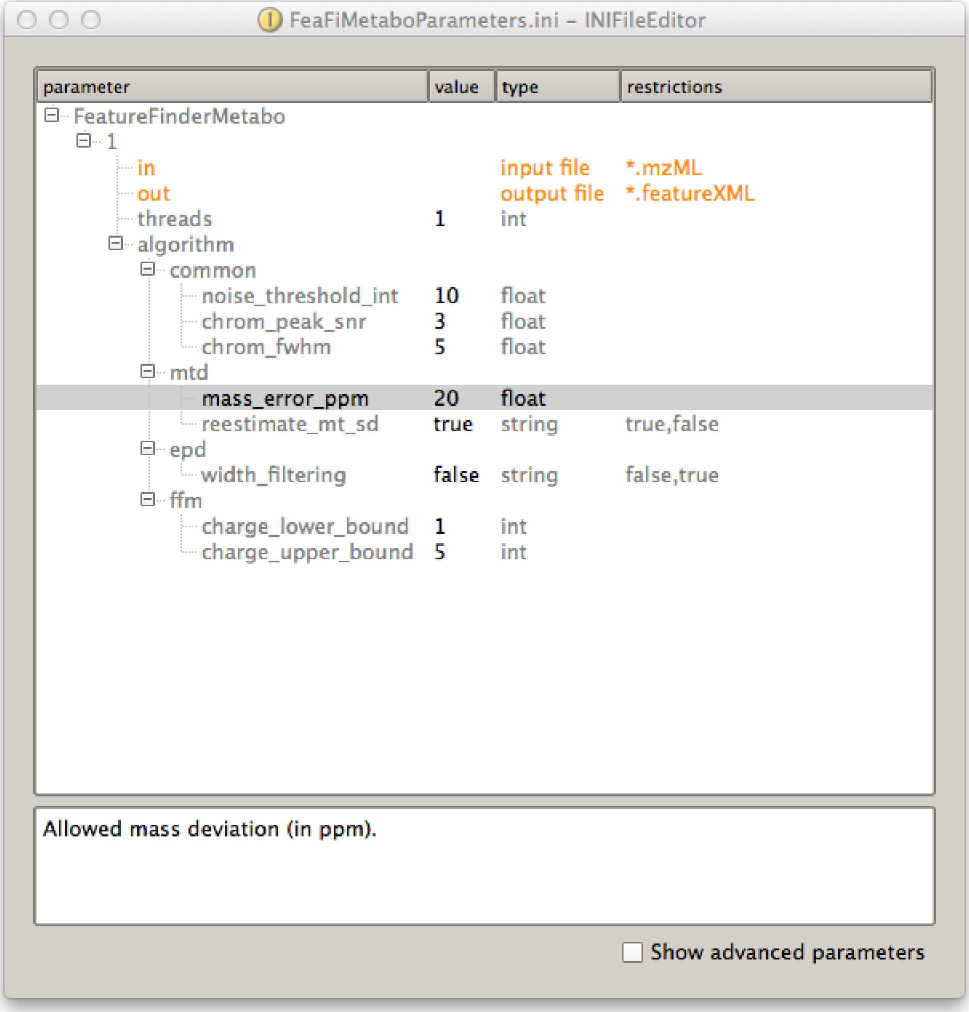


Figure A.6: Feature finding performance with respect to varying settings of the `chrom_fwhm` parameter. Feature detection runs were performed on a dataset with elution profiles simulated with heavy distortion. The `chrom_fwhm` parameter was varied between 1 and 10 s. Since XCMS/CAMERA does not offer a smoothing parameter corresponding to `chrom_fwhm`, we conducted a single run with the parameter *peakwidth* set to the interval $[3, 30]$ seconds, allowing only chromatographic peaks within this range. The set of chromatographic peaks resulting from our algorithm were also filtered by this criterion [53].


FeaFiMetaboParameters.ini – INIFileEditor
□ Show advanced parameters

parameter	value	type	restrictions
⊟ FeatureFinderMetabo			
⊟ 1			
in		input file	*.mzML
out		output file	*.featureXML
threads	1	int	
⊟ algorithm			
⊟ common			
noise_threshold_int	10	float	
chrom_peak_snr	3	float	
chrom_fwhm	5	float	
⊟ mtd			
mass_error_ppm	20	float	
reestimate_mt_sd	true	string	true,false
⊟ epd			
width_filtering	false	string	false,true
⊟ ffm			
charge_lower_bound	1	int	
charge_upper_bound	5	int	

Allowed mass deviation (in ppm).

Figure A.7: Snapshot of the `INIFileEditor` window. The parameters for the `FeatureFinderMetabo` tool can be easily changed with the `INIFileEditor` tool of OpenMS [53].

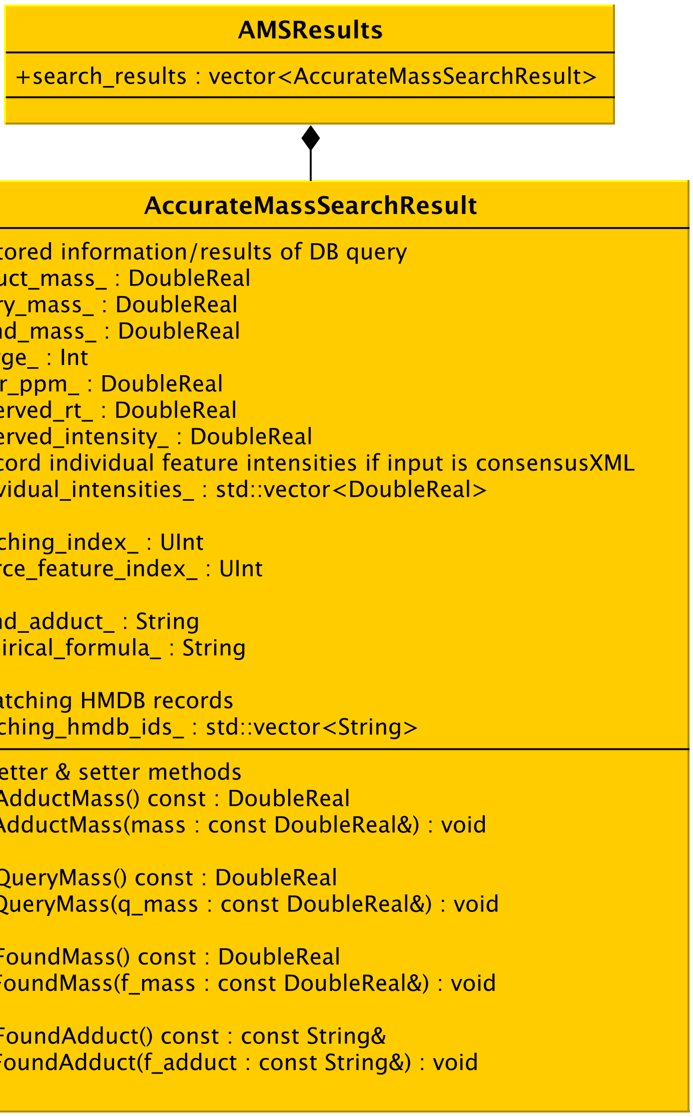

AMSResults
+search_results : vector<AccurateMassSearchResult>
AccurateMassSearchResult
/// Stored information/results of DB query
-adduct_mass_ : DoubleReal
-query_mass_ : DoubleReal
-found_mass_ : DoubleReal
-charge_ : Int
-error_ppm_ : DoubleReal
-observed_rt_ : DoubleReal
-observed_intensity_ : DoubleReal
// record individual feature intensities if input is consensusXML
-individual_intensities_ : std::vector<DoubleReal>
-matching_index_ : UInt
-source_feature_index_ : UInt
-found_adduct_ : String
-empirical_formula_ : String
// matching HMDB records
-matching_hmdb_ids_ : std::vector<String>
/// getter & setter methods
+getAdductMass() const : DoubleReal
+setAdductMass(mass : const DoubleReal&) : void
+getQueryMass() const : DoubleReal
+setQueryMass(q_mass : const DoubleReal&) : void
+getFoundMass() const : DoubleReal
+setFoundMass(f_mass : const DoubleReal&) : void
...
+getFoundAdduct() const : const String&
+setFoundAdduct(f_adduct : const String&) : void
...

Figure A.8: UML diagram of the `AccurateMassSearchResult` class. Instances hereof are created by the AMS algorithm to temporarily store query hits.

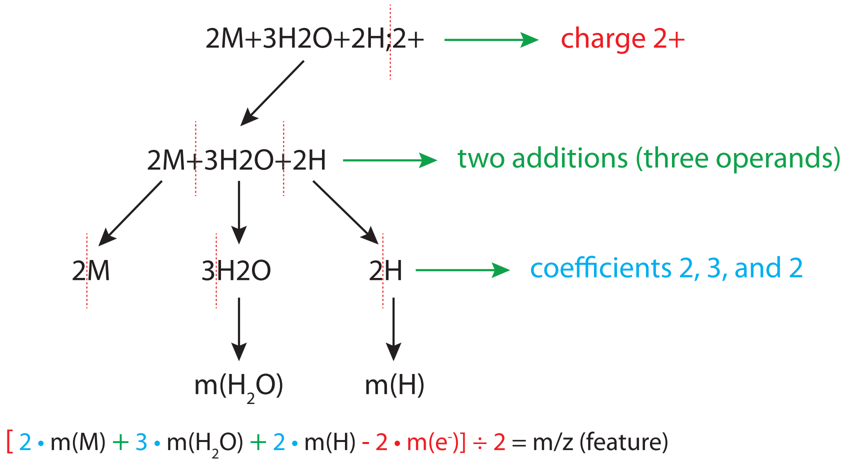

2M+3H2O+2H;2+
charge 2+
2M+3H2O+2H
two additions (three operands)
2M
3H2O
2H
coefficients 2, 3, and 2
m(H2O)
m(H)
[2 • m(M) + 3 • m(H2O) + 2 • m(H) - 2 • m(e-)] ÷ 2 = m/z (feature)

Figure A.9: Parsing of pseudo-molecular formulations in the AMS tool. Based on the adduct string, the parser first removes the suffix beginning with the semicolon and thus determines the adduct charge. Then, the preceding substring is split into a sequence of subcomponents with respect to the "+" and "-" operators. After the stoichiometric coefficients has been extracted, each substring is parsed by an `EmpiricalFormula` object and its monoisotopic mass is calculated [134]. Finally, solving the final equation by $m(M)$ results in the query mass.

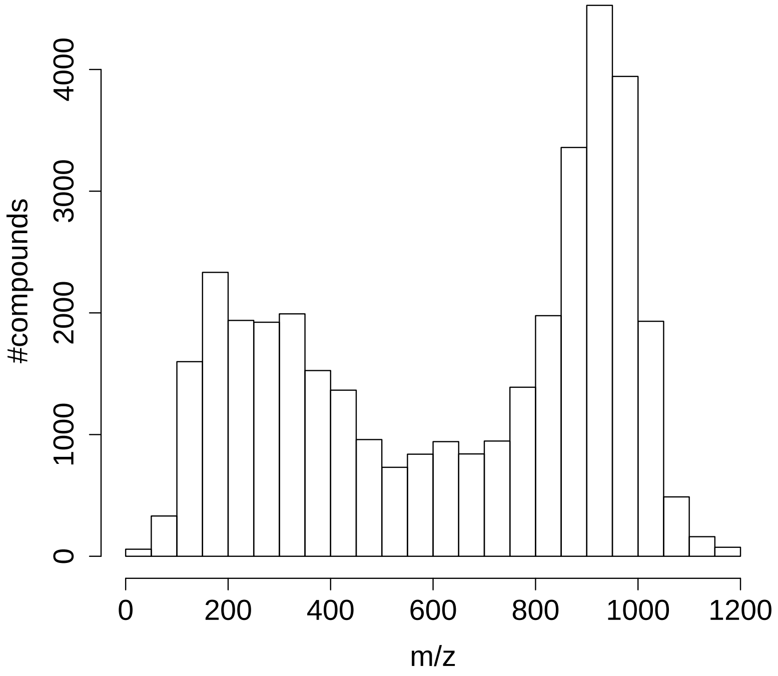


Figure A.10: Distribution of HMDB compound masses below 1,200 Da.

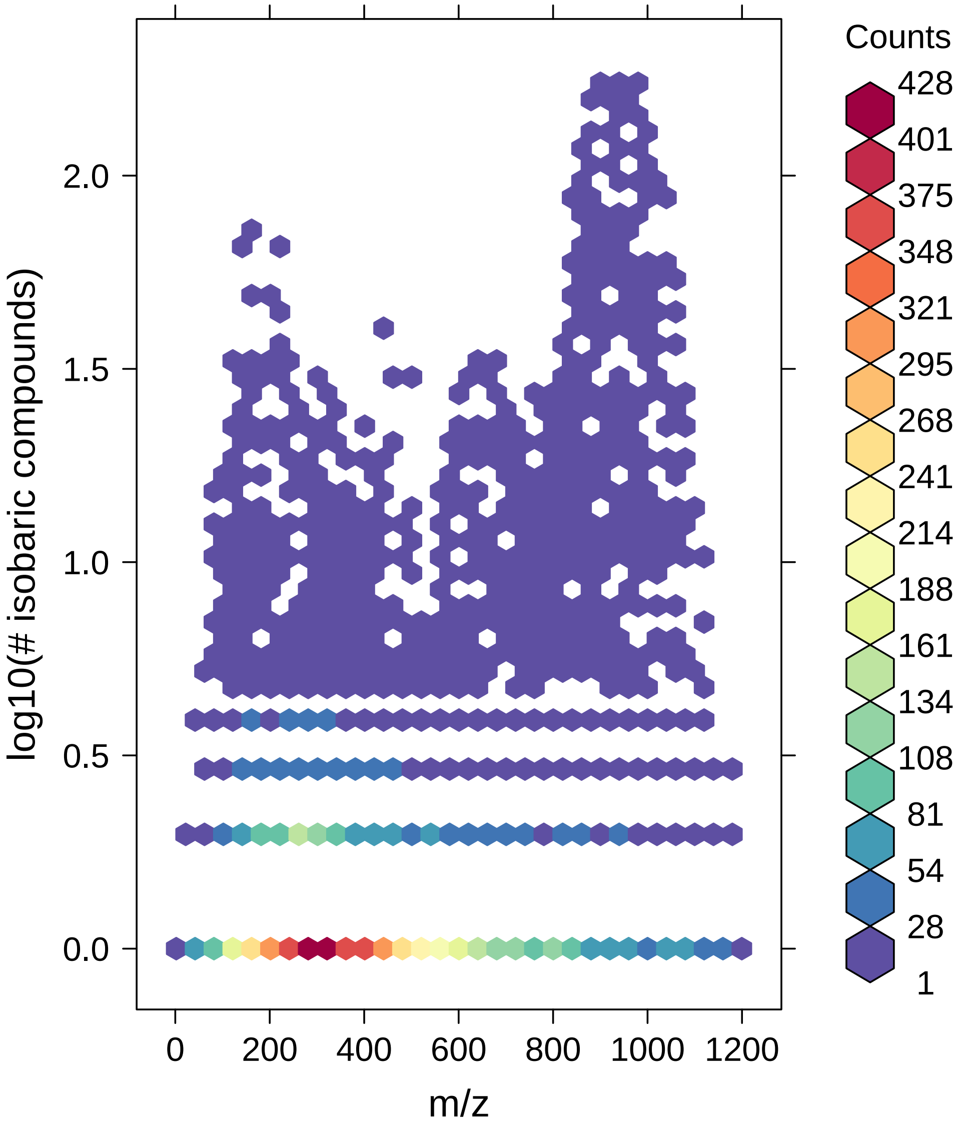
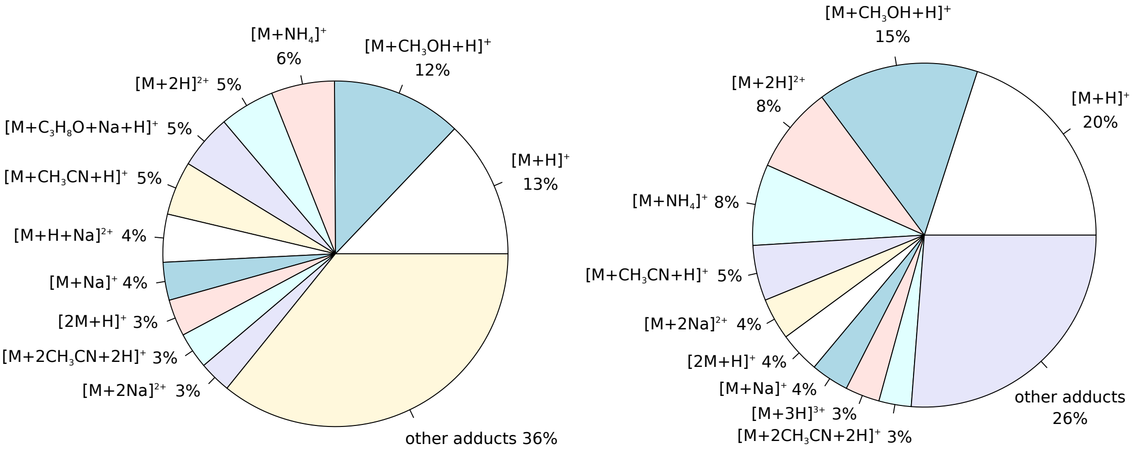


Figure A.11: Two-dimensional histogram of HMDB compound masses. In this *hexagonal binning* plot [197], the frequencies of compounds with identical masses are visualized with respect to m/z.



(a) adducts found with $|\delta m_{\mathrm{AMS}}| \leq 5\,\mathrm{ppm}$ **(b)** adducts found with $|\delta m_{\mathrm{AMS}}| \leq 1\,\mathrm{ppm}$

Figure A.12: Distribution of positive adduct types with respect to mass error tolerance $|\delta m_{\mathrm{AMS}}|$. The underlying numbers were taken from the adduct statistics written out by the AMS tool and correspond to the proportions of features annotated with the respective adduct types. Fractions below 3 % were summarized under the *other adducts* section.

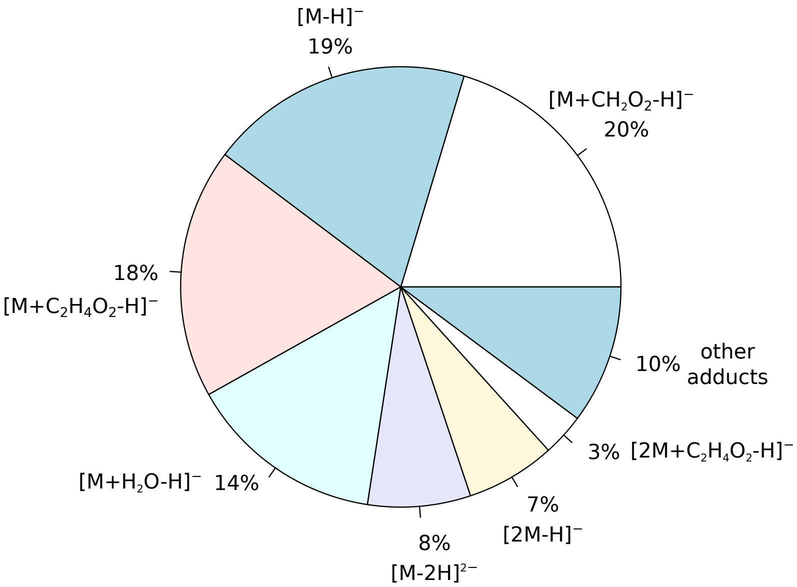


Figure A.13: Distribution of negative adduct types with mass error tolerance $|\delta m_{\mathrm{AMS}}| \leq 5\,\mathrm{ppm}$. Fractions below 3 % were summarized under the *other adducts* section.

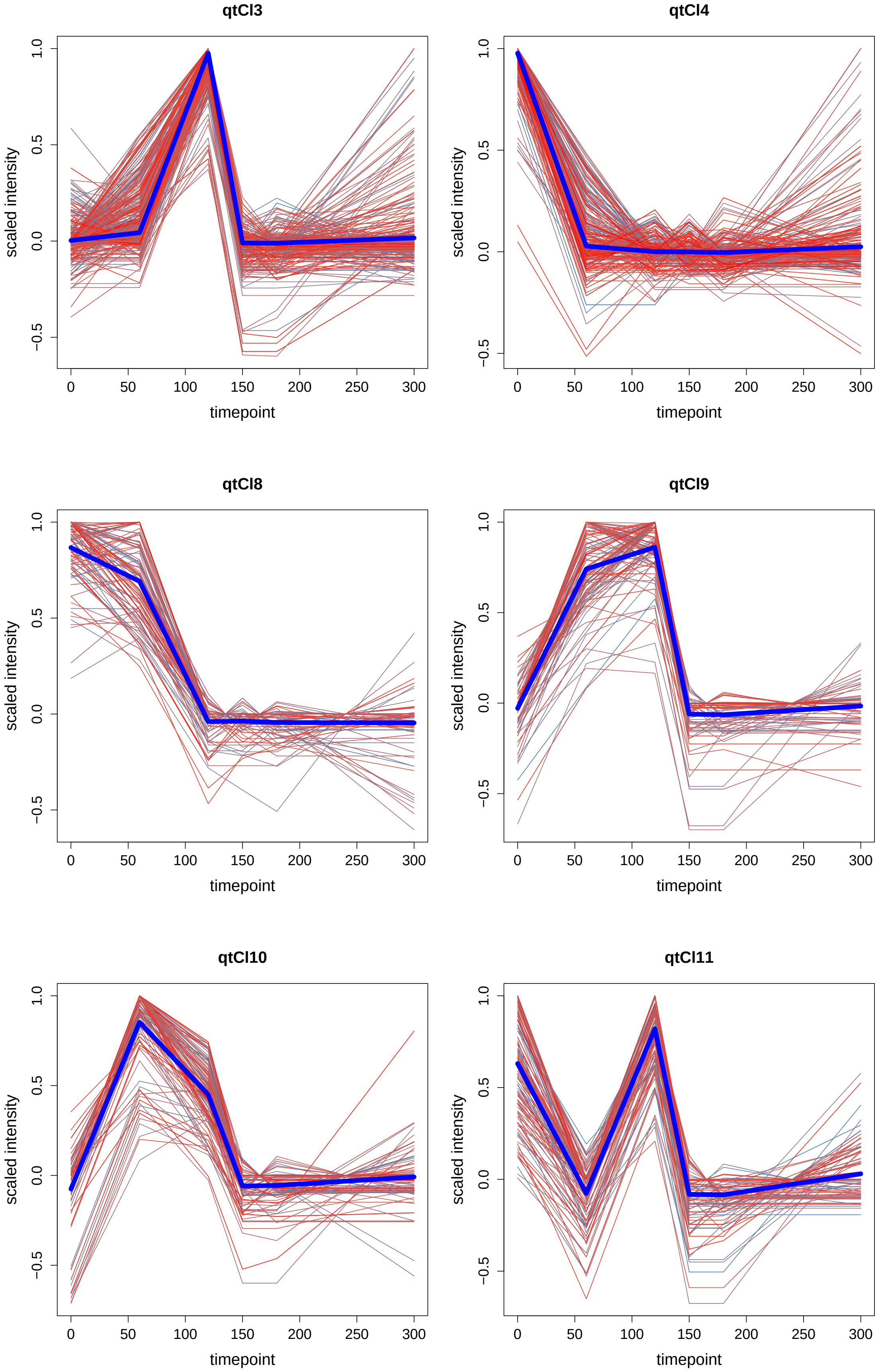


Figure A.14: Clusters representative for the exercise phase (part 1).

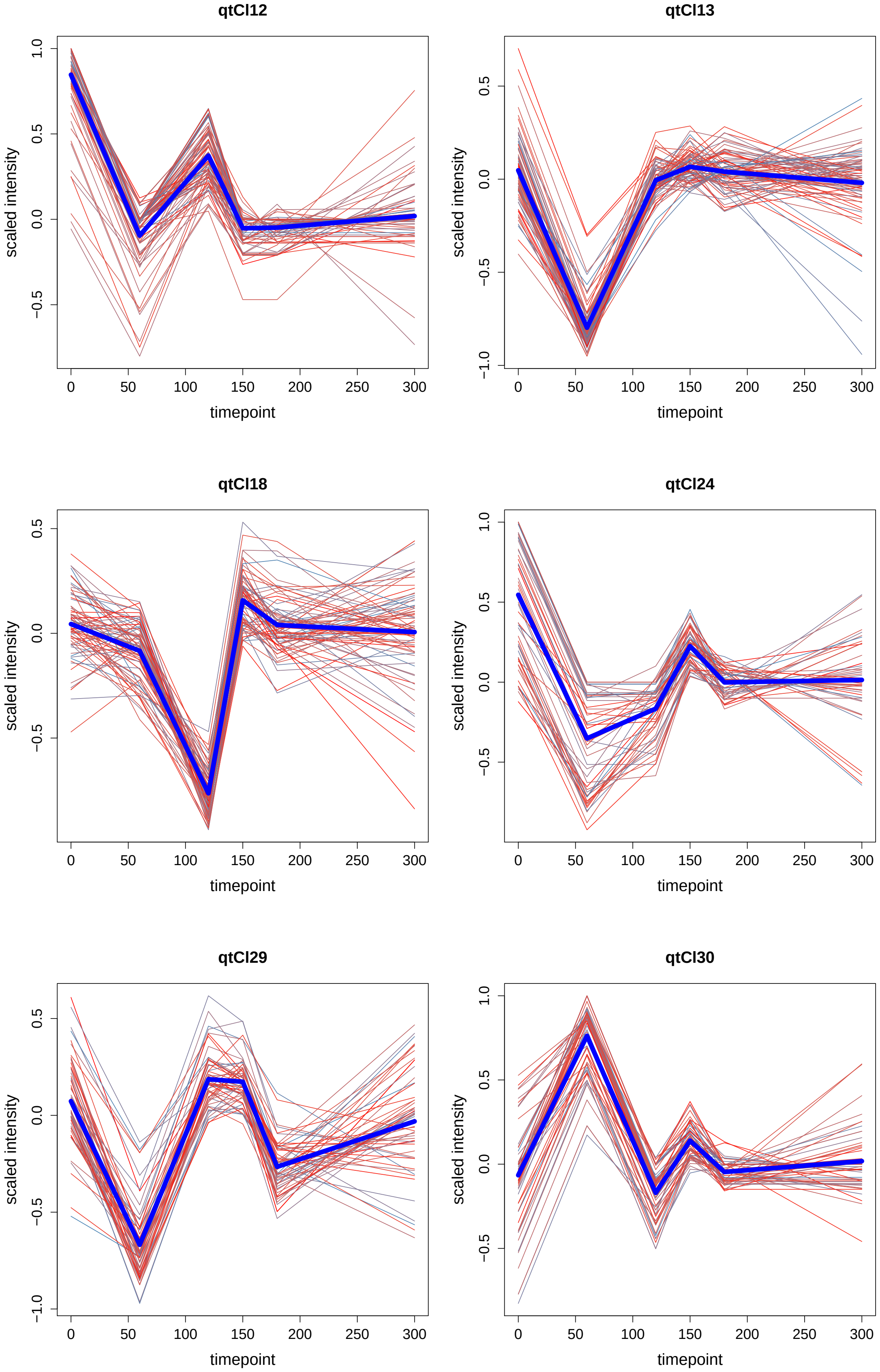


Figure A.15: Clusters representative for the exercise phase (part 2).

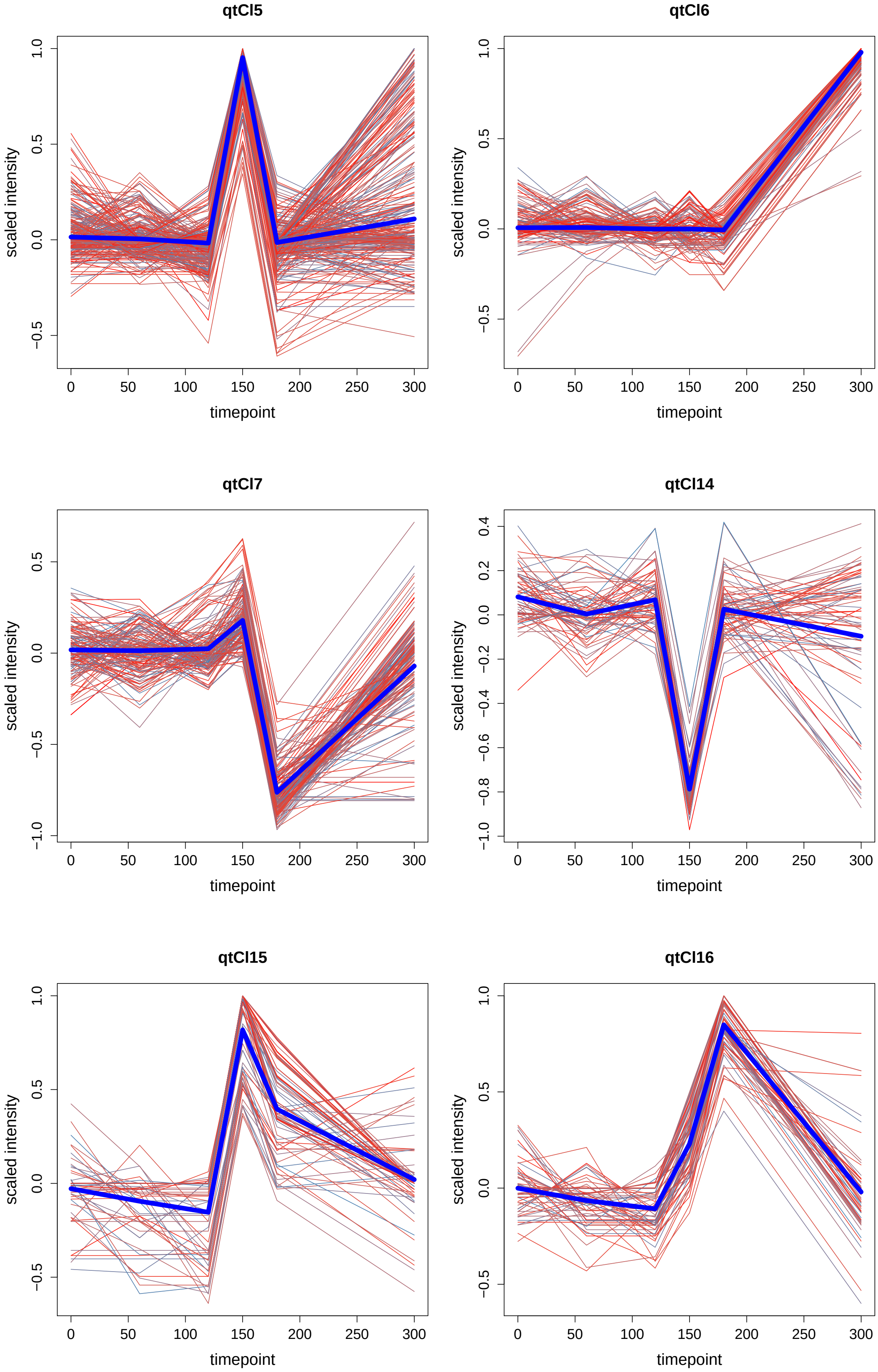


Figure A.16: Clusters representative for the recovery phase (part 1).

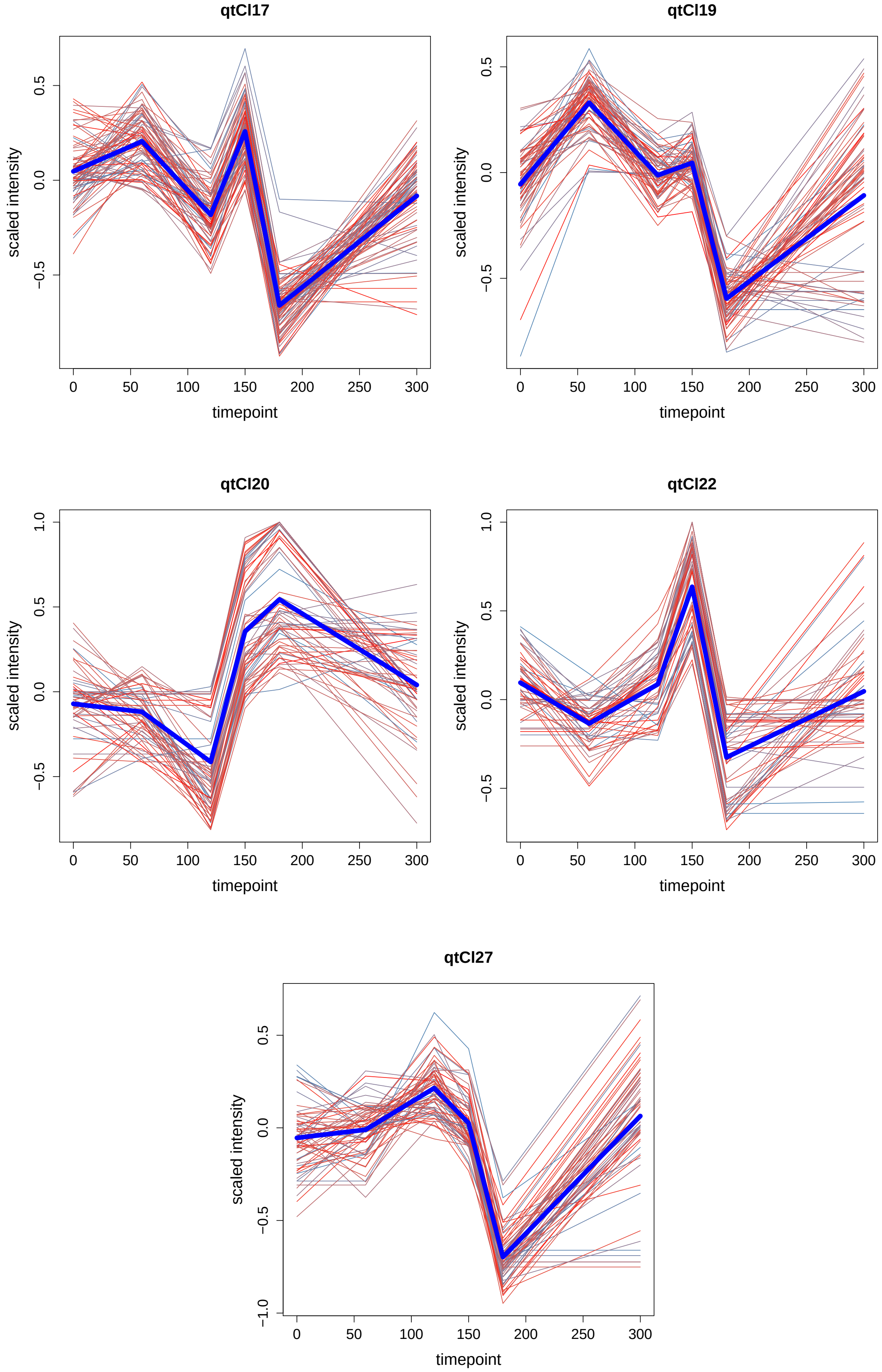


Figure A.17: Clusters representative for the recovery phase (part 2).

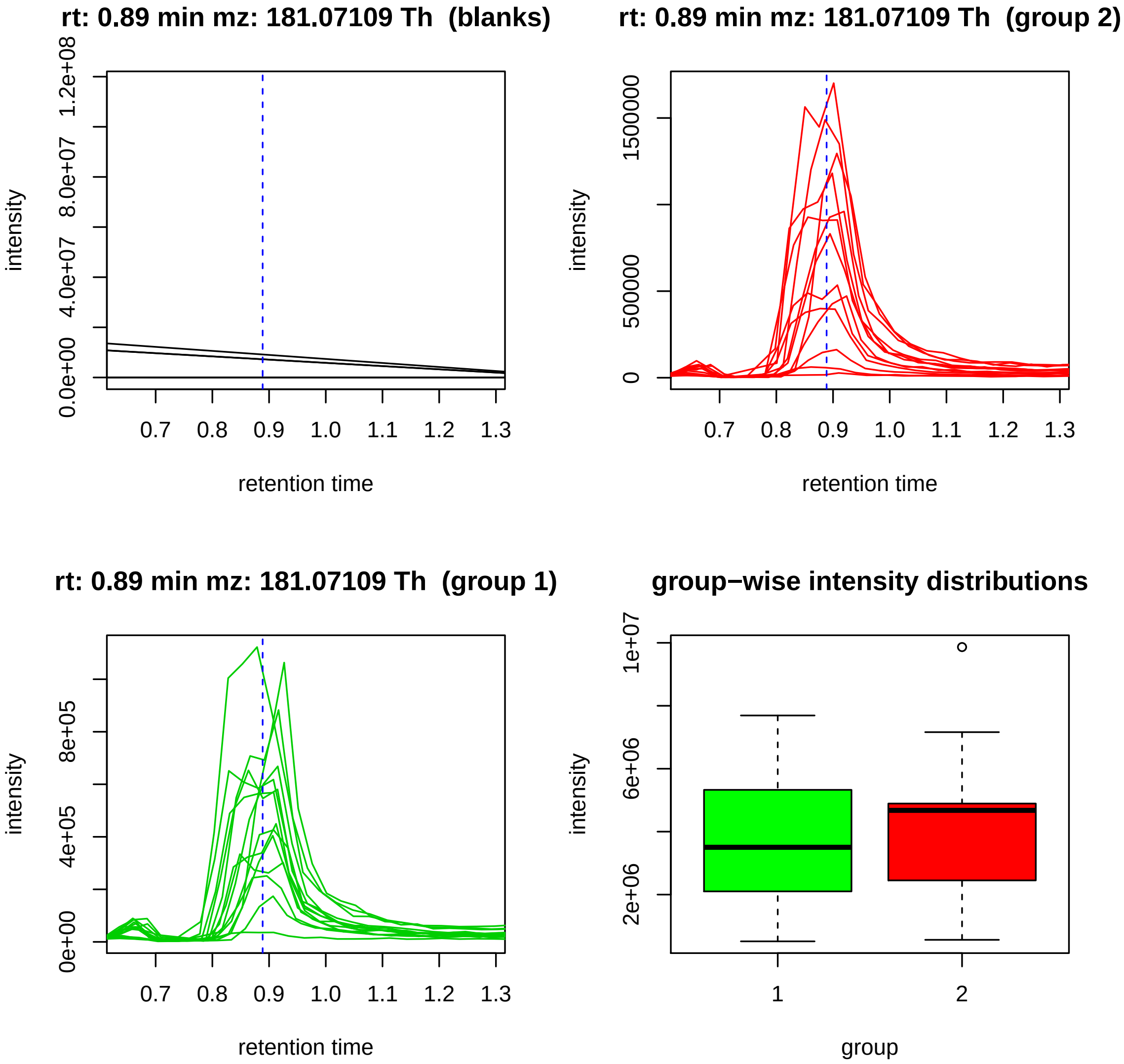


Figure A.18: Profile plots of an up-regulated candidate detected in positive mode. When compared between group 1 (non-risk allele) and group 2 (risk allele), the elution peaks differed significantly in their intensity distributions according to our RankProd analysis, however, could not be distinguished by means of a t-test. Our AMS identification run suggested *D-glucose* among other hexose sugars as potential metabolite IDs (Table A.12 (5)).

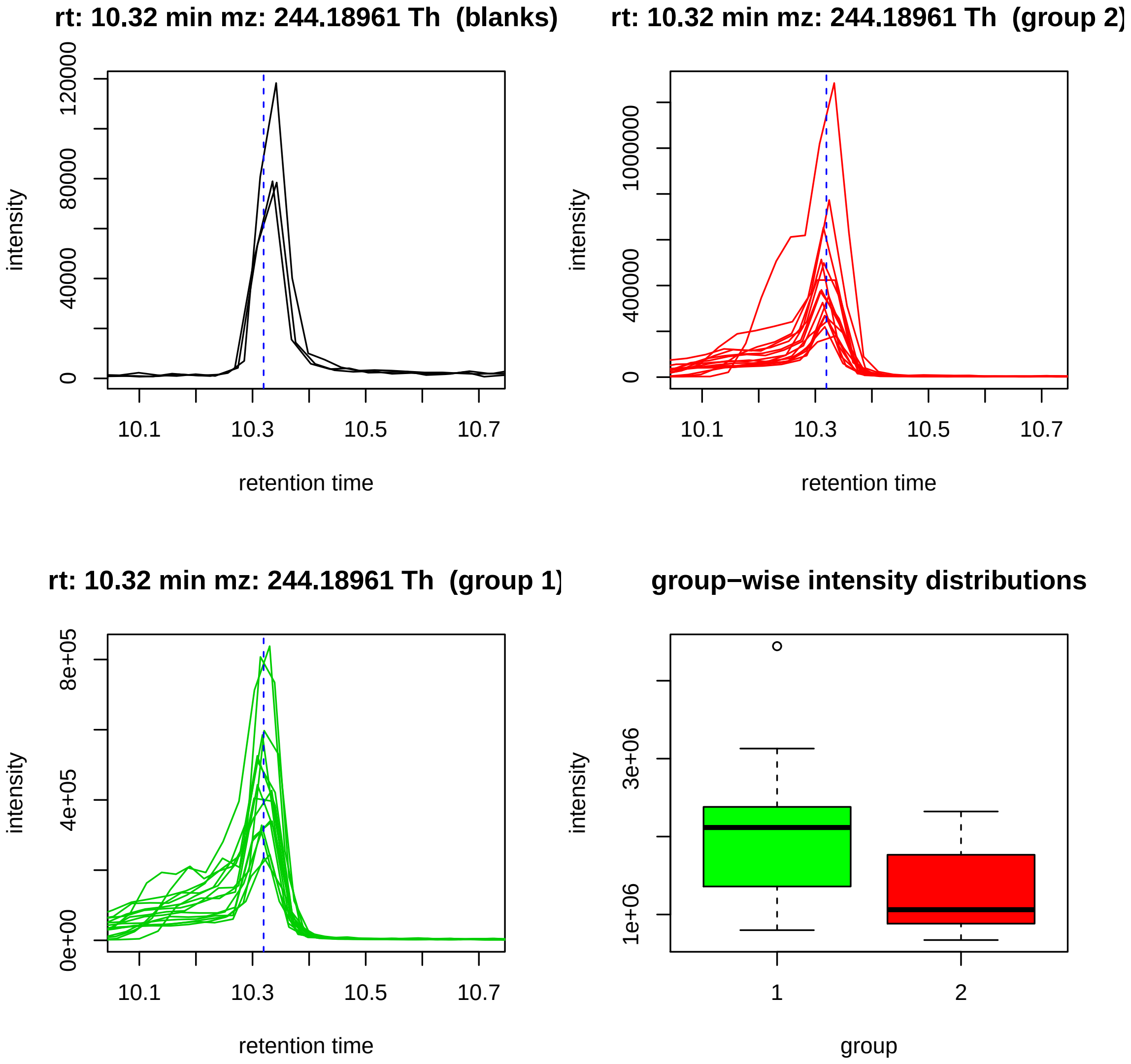


Figure A.19: Profile plots of a down-regulated candidate detected in positive mode. When compared between group 1 (non-risk allele) and group 2 (risk allele), the elution peaks differed significantly in their intensity distributions according to our RankProd analysis and also to a t-test (p-value $= 0.021$, $\alpha = 0.05$). Our AMS identification run suggested *N-undecanoylglycine* as a potential metabolite ID (Table A.14 (3)).

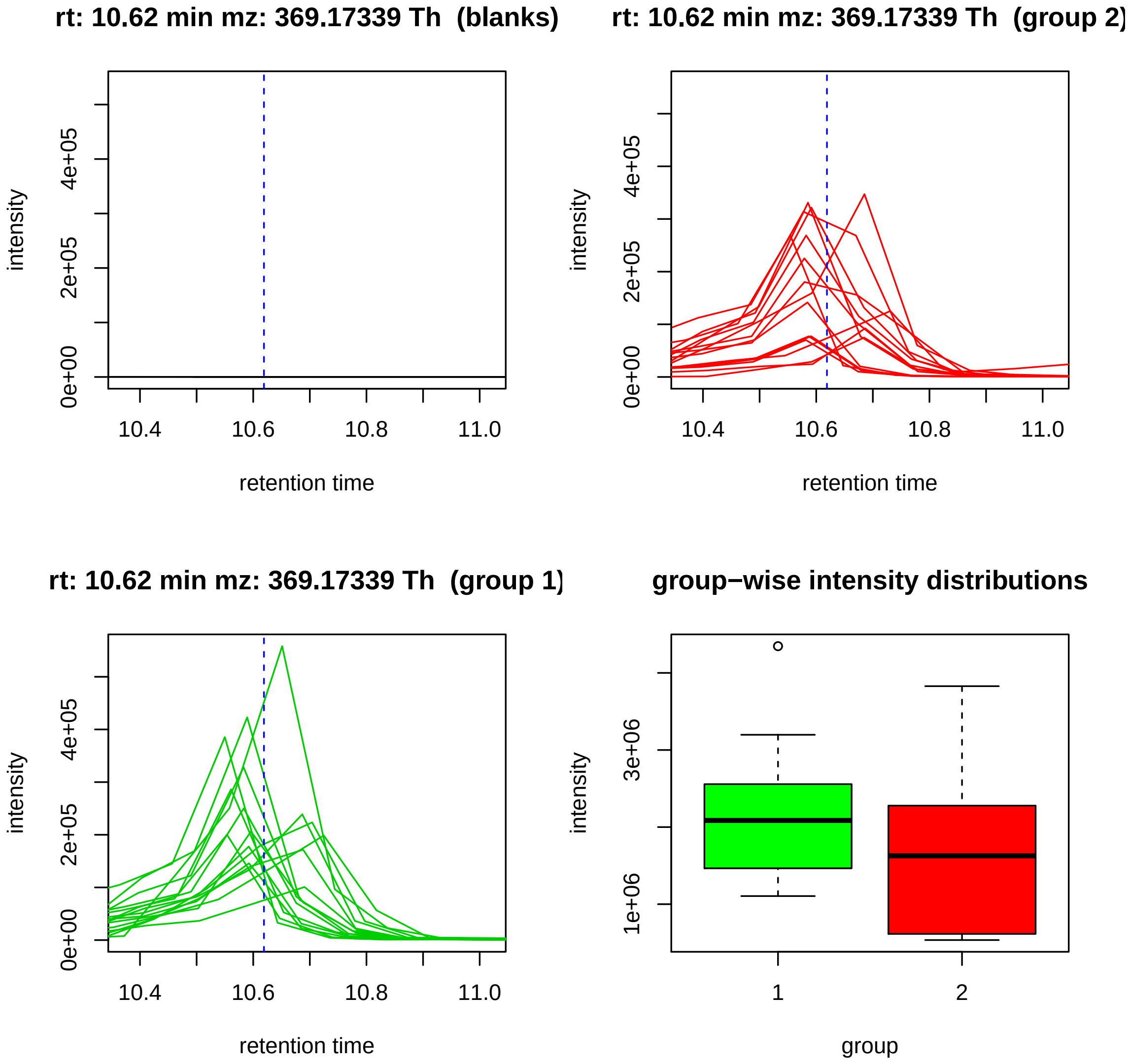


Figure A.20: Profile plots of a down-regulated candidate detected in negative mode. When compared between group 1 (non-risk allele) and group 2 (risk allele), the elution peaks differed significantly in their intensity distributions according to our RankProd analysis, however, could not be distinguished by means of a Wilcoxon rank sum test ($\alpha = 0.05$). Our AMS identification run suggested *androsterone* and *DHT* as potential metabolite IDs (Table A.18 (4).

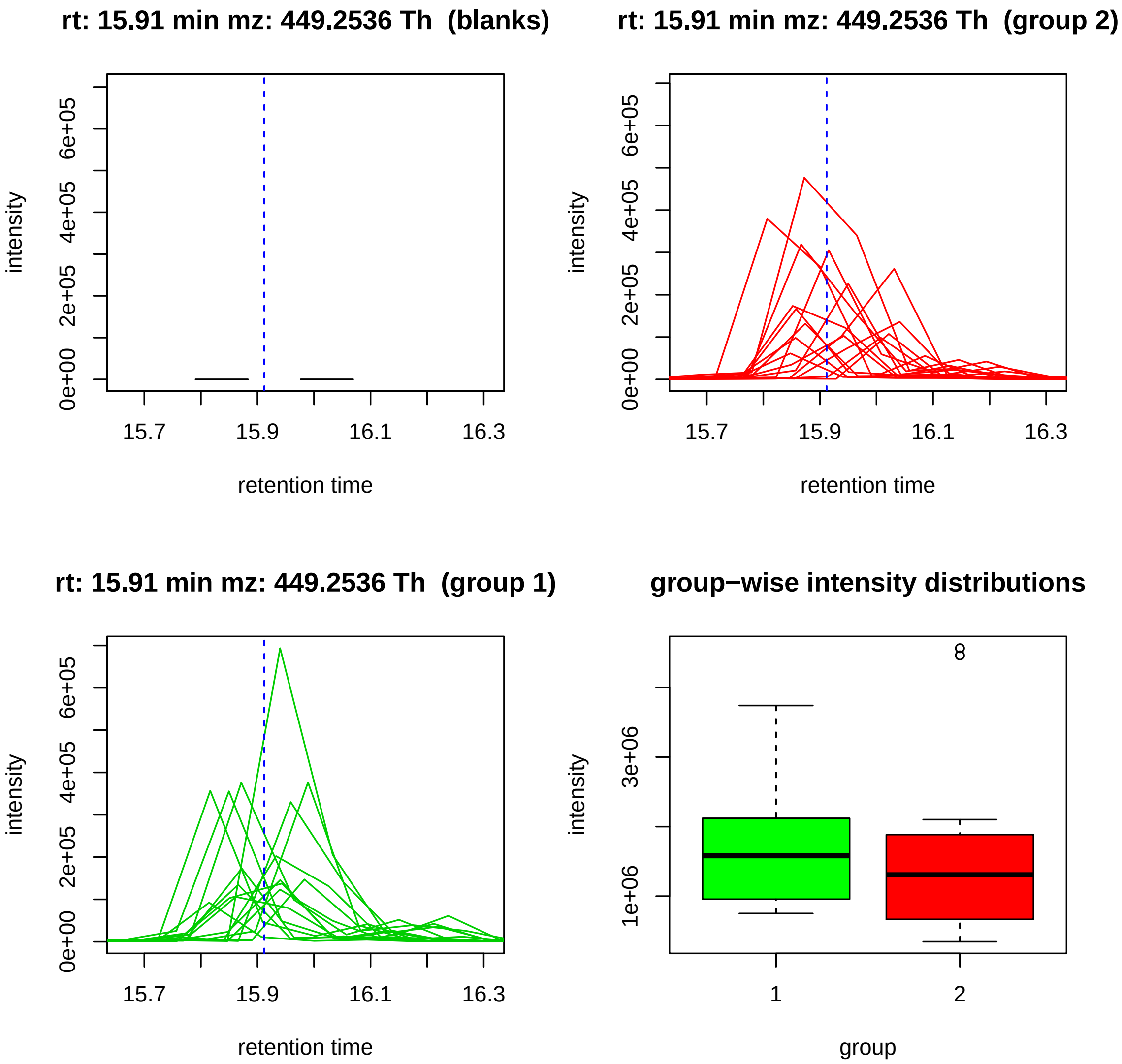


Figure A.21: Profile plots of a down-regulated candidate detected in negative mode. When compared between group 1 (non-risk allele) and group 2 (risk allele), the elution peaks differed significantly in their intensity distributions according to our RankProd analysis, however, could not be distinguished by means of a Wilcoxon rank sum test ($\alpha = 0.05$). Our AMS identification run suggested *androstanediol glucuronide* as a potential metabolite ID Table A.18 (9).

A.3 Tables

Detection parameters	Positive ESI
Capillary voltage (kV)	3.1
Sample cone (V)	30.0
Extraction cone (V)	4.0
Source temperature (°C)	120
Desolvation temperature (°C)	300
Desolvation flow ($\mathrm{L\,h^{-1}}$)	800
Cone ($\mathrm{L\,h^{-1}}$)	50
Detector (kV)	1.8

Table A.1: Configuration of detection parameters for the MS measurements conducted in positive ESI mode [53].

`MSSimulator` parameter	value
`RT:total_gradient_time` [s]	1,500
`RT:sampling_rate` [s]	0.25
`RT:scan_window:min` [s]	0
`RT:scan_window:max` [s]	1500
`RawSignal:resolution:value`	20,000
`RawSignal:resolution:type`	constant
`RawSignal:variation:intensity:scale`	1
`Ionization:mz:lower_measurement_limit` [Da]	100
`Ionization:mz:upper_measurement_limit` [Da]	1,000
detector noise simulation:	
`RawSignal:noise:detector:stddev` [counts]	0, 3, 5, 10, 12, 15, 20, 25, 30, 40, 50, 70 and 100
m/z variation:	
`RawSignal:variation:mz:error_stddev` [ppm]	0, 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50
elution profile distortion:	
`RT:column_condition:distortion`	0 and 10

Table A.2: Parameter configuration of `MSSimulator`. The first nine parameters were common to all simulated datasets. The last three parameters were specific for the simulations of detector noise, m/z variation, and elution profile distortion, respectively [53].

Compound	C3-Carnitine	Nialamide	Sulfadimethoxine	Reserpine	Terfenadine	C16-Carnitine	C18-Carnitine
Correlation	0.9994	0.9957	0.987	0.9987	0.9994	0.9975	0.9803
Concentration $[mg/L]$			intensity mean ± standard deviation				
0	not detected	not detected	not detected	not detected	not detected	not detected	not detected
0.01	not detected	6.3×10^1	2.5×10^2	8.0×10^1	3.8×10^2	2.2×10^2	not detected
		± 0.0 (*)	$\pm 2.7\times10^1$	± 8.3	$\pm 4.3\times10^1$	± 0.0 (*)	not detected
0.02	not detected	1.1×10^2	4.8×10^2	1.4×10^2	6.4×10^2	3.6×10^2	4.5×10^2
		$\pm 1.1\times10^1$	$\pm 7.2\times10^1$	± 9.3	$\pm 8.4\times10^1$	± 3.4	± 7.3
0.05	not detected	2.5×10^2	1.2×10^3	3.6×10^2	1.8×10^3	6.9×10^2	9.1×10^2
		$\pm 1.1\times10^1$	$\pm 9.7\times10^1$	$\pm 3.8\times10^1$	$\pm 1.5\times10^2$	± 8.8	$\pm 1.9\times10^1$
0.1	not detected	2.0×10^2	2.3×10^3	8.3×10^2	3.6×10^3	2.5×10^3	3.2×10^3
		$\pm 1.2\times10^2$ (**)	$\pm 9.5\times10^1$	$\pm 2.5\times10^1$	$\pm 1.3\times10^2$	$\pm 1.9\times10^1$	$\pm 1.8\times10^2$
0.2	not detected	1.0×10^3	3.8×10^3	2.1×10^3	7.7×10^3	6.4×10^3	8.0×10^3
		$\pm 9.9\times10^1$	$\pm 3.3\times10^2$	$\pm 6.7\times10^1$	$\pm 2.7\times10^2$	$\pm 4.7\times10^1$	$\pm 1.2\times10^2$
0.5	not detected	2.5×10^3	8.7×10^3	3.9×10^3	1.4×10^4	1.3×10^4	1.8×10^4
		$\pm 1.1\times10^2$	$\pm 7.8\times10^2$	$\pm 3.0\times10^2$	$\pm 5.0\times10^2$	$\pm 1.0\times10^2$	$\pm 2.1\times10^2$
1	3.1×10^2	4.5×10^3	1.7×10^4	8.1×10^3	2.5×10^4	4.8×10^4	5.6×10^4
	$\pm 1.4\times10^1$	$\pm 1.3\times10^2$	$\pm 1.2\times10^3$	$\pm 7.2\times10^2$	$\pm 1.5\times10^3$	$\pm 1.4\times10^2$	$\pm 5.1\times10^2$
2	5.3×10^2	9.7×10^3	3.4×10^4	1.9×10^4	5.5×10^4	6.9×10^4	9.3×10^4
	$\pm 2.1\times10^1$	$\pm 1.1\times10^3$	$\pm 1.5\times10^3$	$\pm 1.2\times10^3$	$\pm 1.3\times10^3$	$\pm 5.9\times10^2$	$\pm 5.7\times10^3$
5	1.1×10^3	2.2×10^4	7.2×10^4	5.7×10^4	1.4×10^5	1.5×10^5	2.3×10^5
	$\pm 1.0\times10^2$	$\pm 2.2\times10^3$	$\pm 4.1\times10^2$	$\pm 2.3\times10^3$	$\pm 1.8\times10^3$	$\pm 2.3\times10^3$	$\pm 6.4\times10^3$
10	2.0×10^3	3.7×10^4	1.1×10^5	1.0×10^5	2.6×10^5	3.2×10^5	3.2×10^5
	$\pm 1.3\times10^2$	$\pm 1.4\times10^3$	$\pm 4.0\times10^3$	$\pm 1.1\times10^3$	$\pm 4.0\times10^3$	$\pm 1.2\times10^3$	$\pm 2.0\times10^3$

Table A.3: Feature intensities for spiked-in compounds. The first row contains the Pearson correlations between the concentrations and the corresponding feature intensities for each standard compound. (*) detected in one of three, (**) detected in two of three replicates [53].

neutral mass	empirical formula	HMDB ID 1	HMDB ID 2	HMDB ID 3	...
...	...	...	...	...	
136.03467	C8H8S	HMDB31825	n.a.	n.a.	
136.03717	C4H8O5	HMDB00613	HMDB00943	n.a.	
136.03804	C5H12S2	HMDB31160	HMDB33053	HMDB33054	
...	...	...	..	...	

Table A.4: Structure of the AMS tool's tabular HMDB mapping file. Each row contains a unique neutral mass/empirical formula that is extended by all matching HMDB metabolites. The complete mapping is stored as `HMDBMappingFile.tsv` in the OpenMS file hierarchy.

MTD mzTab-version 1.0.0

SMH	identifier	unit_id	chemical_formula	smiles
SML	HMDB00125	null	C10H17N3O6S	N[C@@H](CCC(=O)N...
...	...	...	...	...

inchi_key	description	mass_to_charge	charge	retention_time	...
InChI=1S/...	Glutathione	308.09012	1	76.37	...
...	...	...	...	...	...

opt_ppm_error	opt_adduct_ion	opt_id_group
-2.86	M+H;1+	1042
...	...	...

Table A.5: Example for an `mzTab` output file. The identification result of glutathione is stored as an individual row in an `mzTab` report file. While the most essential columns are shown, some columns required by the standard were left out for the sake of clarity.

```
ACCESSION: KNA00121
RECORD_TITLE: L-Glutathione (reduced form); LC-ESI-ITFT; MS; POS
DATE: 2011.08.03 (Created 2009.11.17)
AUTHORS: Takahashi H, Kanaya S, Ogasawara N, Graduate School of
Information Science, NAIST
LICENSE: CC BY-SA
CH$NAME: Glutathione
CH$NAME: 5-L-Glutamyl-L-cysteinylglycine
CH$NAME: N-(N-gamma-L-Glutamyl-L-cysteinyl)glycine
CH$NAME: gamma-L-Glutamyl-L-cysteinyl-glycine
CH$NAME: GSH
CH$NAME: Reduced glutathione
CH$NAME: L-Glutathione (reduced form)
CH$COMPOUND_CLASS: Natural Product
CH$FORMULA: C10H17N3O6S
CH$EXACT_MASS: 307.08381
CH$SMILES: OC(=O)CNC(=O)C(CS)NC(=O)CCC(N)C(O)=O
CH$IUPAC: InChI=1S/C10H17N3O6S/c11-5(10(18)19)1-2-7(14)13-6(4-20)9(17) ...
CH$LINK: CAS 70-18-8
CH$LINK: CHEBI 16856
CH$LINK: KEGG C00051
CH$LINK: KNAPSACK C00001518
CH$LINK: NIKKAJI J10.686K
CH$LINK: PUBCHEM 3353
AC$INSTRUMENT: LTQ Orbitrap XL, Thermo Scientfic
AC$INSTRUMENT_TYPE: LC-ESI-ITFT
AC$MASS_SPECTROMETRY: MS_TYPE MS
AC$MASS_SPECTROMETRY: ION_MODE POSITIVE
AC$MASS_SPECTROMETRY: COLLISION_ENERGY 35eV
AC$MASS_SPECTROMETRY: IONIZATION ESI
AC$CHROMATOGRAPHY: COLUMN_NAME TOSOH TSKgel ODS-100V  5um Part no. 21456
AC$CHROMATOGRAPHY: FLOW_GRADIENT 0min:3%, 45min:97%, 50min:97%, 50.1:3%, ...
AC$CHROMATOGRAPHY: FLOW_RATE 0.5 ml/min
AC$CHROMATOGRAPHY: RETENTION_TIME 8.858290 min
AC$CHROMATOGRAPHY: SOLVENT 0.1%formate-water / 0.1%formate-acetonitrile
MS$FOCUSED_ION: BASE_PEAK 308.091064
PK$NUM_PEAK: 87
PK$PEAK: m/z int. rel.int.
  ...
  224.127985 2952810.250000 27
  ...
  290.080490 2041832.000000 18
  ...
  308.090822 110826864.000000 999
  ...
  309.093254 14585537.000000 131
  ...
  310.087177 4548931.000000 41
  ...
  615.175264 6018219.500000 54
  ...

//
```

Table A.6: Abridged content of the KNA00121 record from MassBank. This record contains the MS/MS spectrum of reduced glutathione as a peak list and information about its acquisition.

Table A.7: Validation dataset for our metabolite ID strategies. The dataset comprises 54 authentic standard compounds ranging from all amino acids, TCA metabolites, and more common compounds. Notes: (1) This ID was further confirmed by the presence of cystine (matches MassBank spectrum PB000446 with a score of 24.74 and an E-value of $e^{-4.56}$) (2) Leucine was ranked forth since MSM matched isoleucine, norleucine, and alloisoleucine with slightly higher scores. (3) This potential ammonium adduct of fumarate coincided with the more likely proton adduct of aspartic acid and thus was discarded from the list of candidate IDs. (4) This adduct collided with the more likely deprotonated citric/isocitric acid and thus was discarded. (5) Retinoic acid shared the first AMS rank with numerous other isobaric candidates. (6) We could not observe a clear metabolite feature, however, the MSM match indicated the presence of an $[M+H]^+$ adduct. (7) The retention time of the matched feature seems unlikely (too high). (8) Retinal itself was not detected but its oxidized form (4-OH-retinal).

Authentic standard	Detected adduct(s)	Suggested ID		ID rank		mass error	MSM score	MSM E-value	Note
		AMS	MSM	AMS	MSM	[ppm]		exponent	
Amino acids									
Arginine	$[M+H]^+$	HMDB00517	PB000419	1	1	0.19	18.66	-21.13	
Asparagine	$[M+H]^+$	HMDB00168	PB000458	2	1	0.95	17.93	-8.43	
Aspartic acid	$[M+H]^+$	HMDB00191	PB000453	1	1	1.07	11.43	0.51	
	$[M-H]^-$	HMDB00191	CE000465	1	1	1.6	11.10	NA	
Cysteine	$[M+H]^+$	HMDB00574	PB000438	1	1	1.90	11.27	-6.21	see (1)
Glutamine	$[M+H]^+$	HMDB00641	PB000467	1	1	0.06	14.42	-6.91	
	$[M-H]^-$	HMDB00641	CE000298	1	1	2.01	12.15	NA	
Glutamic acid	$[M+H]^+$	HMDB00148	PB000463	1	1	0.03	12.49	-10.16	
	$[M-H]^-$	HMDB00148	BML01298	1	1	1.83	10.14	NA	
Histidine	$[M+H]^+$	HMDB00177	PB000423	1	1	0.01	17.75	-23.34	
	$[M-H]^-$	HMDB00177	PB000423	2	1	2.2	17.86	NA	
Isoleucine	$[M+H]^+$	HMDB00172	PB000397	1	1	1.36	9.95	-6.74	
Leucine	$[M+H]^+$	HMDB00687	PB000393	1	4	1.24	9.90	-9.12	see (2)

continued on next page

Table A.7 – continued from previous page

Authentic standard	Detected adduct(s)	Suggested ID		ID rank		mass error	MSM score	MSM E-value	Note
		AMS	MSM	AMS	MSM	[ppm]		exponent	
Lysine	$[M+H]^+$	HMDB00182	PB000428	1	1	0.75	15.90	-12.21	
Methionine	$[M+H]^+$	HMDB00696	PB000442	1	1	0.39	18.27	-11.99	
Phenylalanine	$[M+H]^+$	HMDB00159	PB006064	1	1	0.07	9.90	-1.70	
Proline	$[M+H]^+$	HMDB00162	PB000451	2	1	2.92	11.01	-11.19	
Serine	$[M+H]^+$	HMDB00187	PB000400	3	1	3.84	11.03	-7.03	
Threonine	$[M+H]^+$	HMDB00167	PB000405	1	1	2.12	10.32	-3.13	
Tryptophan	$[M+H]^+$	HMDB00929	PB000416	1	1	0.39	17.39	-16.09	
Tyrosine	$[M+H]^+$	HMDB00158	PB000412	1	1	0.18	12.52	-0.87	
Valine	$[M+H]^+$	HMDB00883	PB000389	1	1	2.08	9.91	-11.05	
Alanine	none								
Glycine	none								
TCA metabolites									
Citric acid	$[M-H]^-$	HMDB00094	PR100481	2	1	2.45	12.72	NA	
DL-Isocitric acid	$[M-H]^-$	HMDB00193	PR100481	4	1	2.73	14.16	-16.04	
L-(-)-malic acid	$[M-H]^-$	HMDB00156	PR100541	1	1	1.82	8.88	NA	
Succinate	$[M-H]^-$	HMDB00254	PR101002	2	1	1.23	9.91	NA	
Pyruvate	none		not in DB						
Succinyl coenzyme A	none		not in DB						
α-Ketoglutaric acid	$[M+H]^+$, $[M-H]^-$	HMDB00208	not in DB	1		0.51			
Acetyl coenzyme A	none								

continued on next page

Table A.7 – continued from previous page

Authentic standard	Detected adduct(s)	Suggested ID		ID rank		mass error	MSM score	MSM E-value	Note
		AMS	MSM	AMS	MSM	[ppm]		exponent	
Fumarate	$[M+NH_4]^+$	HMDB00134	not in DB	2		1.23			see (3)
Oxaloacetic acid	$[M+C_2H_4O_2-H]^-$	HMDB00223	not in DB	6		3.97			see (4)
Common compounds									
AMP	$[M+H]^+$	HMDB00045	KNA00199	1	1	0.55	9.90	-5.46	
	$[M-H]^-$	HMDB00045	PR100515	1	1	1.2	11.08	-5.80	
Hydrocortisone	$[M+H]^+$	HMDB14879	CO000223	1	1	1.05	87.68	NA	
NADH	$[M+2H]^{2+}$	HMDB01487	KNA00269	1	1	1.03	11.30	0.30	
	$[M+H]^+$	HMDB01487	not found	1		0.69			
	$[M-H]^-$	HMDB01487	not found	1		2.58			
L-ascorbic acid	$[M+H]^+$	HMDB00044	JP006292	1	1	0.20	17.51	-4.97	
Pyridoxine	$[M-H]^-$	HMDB00239	PR100602	2	1	1.06	10.74	NA	
Imidazole	$[M+CH_3CN+H]+$	HMDB01525	not found	3		4.87			
Retinoic acid	$[M+2H]^{2+}$	HMDB01852	not found	1		-1.21			see (5)
Histamine	$[M+H]^+$	HMDB00870	not found	2		3.42			
PC(16:0/16:0)	none		not in DB						
Uric acid	none	not found	MT000083		1		13.40	-5.15	see (6)
Spermine	$[M+K]^+$	HMDB01256	not found	2		1.76			see (7)
Glucose-6-phosphate	none								
Oleic acid	$[M+H]^+$	HMDB00207	not in DB	1		-0.88			
	$[M+CH_2O_2-H]^-$	HMDB00207	not in DB	2		2.60			

continued on next page

Table A.7 – continued from previous page

Authentic standard	Detected adduct(s)	Suggested ID		ID rank		mass error	MSM score	MSM E-value	Note
		AMS	MSM	AMS	MSM	[ppm]		exponent	
Cholesterol	$[M + CH_3OH + H]^+$	HMDB00067	not found	1		-1.31			
Folic acid	$[M + 3\,CH_3CN + 2\,H]^{2+}$	HMDB00121	not found	1		-0.47			
UDP	$[2M + 3\,H_2O + 2\,H]^{2+}$	HMDB00295	not in DB	2		0.83			
Retinal	none		not in DB						see (8)
NAD+	none		not found						
UTP	none		not in DB						
ATP	none		not in DB						
L-lactic acid	none		not in DB						
ADP	none		not found						
UMP	none		not in DB						
L-glutathione (red.)	none		not found						

positive mode			negative mode		
Compound	rt [min]	XLogP	Compound	rt [min]	XLogP
Carnitine	0.77	-1.74	Uric acid	0.84	-0.71
Acetyl-L-carnitine	0.88	-1.00	Hippuric acid	3.48	0.23
Leucine	0.89	-1.39	GCA	9.09	3.34
Proprionyl-L-carnitine	0.94	-0.75	CA	10.36	4.09
Tryptophan (kynurenate)	2.64	0.66	GCDCA	10.55	5.42
Valeryl-L-carnitine	3.46	0.39	LPE C20:5	12.59	5.48
Xanthurenic acid	4.79	0.03	LPE C18:2	13.16	5.32
Octenoyl-L-carnitine	6.36	2.48	LPC C18:2	13.23	5.76
Octanoyl-L-carnitine	7.49	2.10	LPE C16:0	13.52	5.02
Decanoyl-L-carnitine	9.39	3.23	LPE C20:3	13.79	6.13
Bilirubin	9.69	2.77	LPE C18:1	14.06	5.65
Carnitine C12:1	10.24	4.76	FFA C22:6	17.34	8.83
LPC C18:3	10.52	5.43	FFA C16:1	17.54	7.05
Dodecanoyl-L-carnitine	10.95	4.37	FFA C16:0	18.25	7.57
Acyl-L-carnitine C14:2	10.96	4.67			
Tetradecenoyl-L-carnitine	11.70	5.89			
LPC C16:1	12.62	4.95			
LPC C20:1	14.09	7.22			

Table A.8: Training datasets for the RPLC rt prediction models. Based on this list of confirmed metabolite IDs, linear regression models were built to predict and filter out unlikely IDs from AMS results.

Subject characteristics ($n = 8$)			
Age	yrs	20.9	± 0.6
Weight	kg	71.6	± 3.1
Height	cm	178.2	± 1.8
BMI	kg/m^2	22.6	± 1.0
Systolic BP	mmHg	135.3	± 3.3
Diastolic BP	mmHg	79.1	± 3.0
Hb	mmol/L	9.6	± 0.2
Leucocytes	10^9/L	5.9	± 0.5
Trial performance	W	30.3	± 1.6

Table A.9: Characteristics of the individuals participating in the one-legged knee extensor experiments. All values are reported as mean ± standard error of mean.

	CC (non-risk) ($n = 15$) Mean±SD	TT (rs7903146) ($n = 15$) Mean±SD	p
female/male	8/7	7/8	1
age	50.6 ± 6.3	50.2 ± 9.6	0.7
BMI	28.7 ± 2.3	28.6 ± 3.2	0.72
glucose 0 min	5.5 ± 0.6	5.7 ± 0.8	0.43
glucose 120 min	9.0 ± 1.1	9.1 ± 0.9	0.46
insulin 0 min	63.8 ± 28.4	63.0 ± 32.8	0.71
insulin 120 min	412.4 ± 162.3	403.3 ± 273.6	0.29
C-peptide 0 min	658.2 ± 199.4	597.9 ± 228.8	0.53
C-peptide 120 min	1813.7 ± 495.7	1677.5 ± 707.7	0.27
HOMA-β	102.2 ± 19.6	88.1 ± 20.7	0.04
insulin sensitivity index (Matsuda)	9.4 ± 3.7	10.7 ± 6.0	0.79

Table A.10: Subject characteristics of our TCF7L2 cohort. The subjects were selected such that the number of males and females was balanced and the mean age and mean BMI were comparable between both groups. The p-value indicates if there was a significant difference between the group means (significance level $\alpha = 0.05$).

centroid rt	centroid m/z	fold change	fold change	PFP	p-value	ID
[min]	[Th]	(RankProd)	(medians)			index
16.99	722.50244	1.5	1.5	0.036	0.000	
17.04	634.45024	1.53	1.66	0.024	0.000	
17	678.4764	1.51	1.59	0.023	0.000	
16.97	766.52858	1.46	1.39	0.021	0.000	
17.07	590.42408	1.52	1.57	0.02	0.000	
0.82	144.10113	n/a	0.97	0.039	0	
17.11	546.39798	1.49	1.54	0.075	0.001	
16.95	810.55473	1.41	1.28	0.07	0.001	
15.82	507.32125	1.51	1.42	0.067	0.001	
16.88	898.60714	1.37	1.44	0.083	0.001	
12.92	450.31954	n/a	0.96	0.077	0.001	
0.87	223.09133	n/a	1.33	0.083	0.001	
0.72	190.91106	1.3	1.06	0.088	0.001	
22.9	704.52006	1.42	1.28	0.087	0.001	
10.67	338.0519	n/a	1.8	0.083	0.001	
16.9	854.58087	1.37	1.3	0.081	0.001	
18.33	506.25194	n/a	1.46	0.092	0.002	
0.71	242.92423	1.2	1.04	0.108	0.002	
14.6	476.27548	n/a	1.31	0.103	0.002	1
17.14	502.3722	1.44	1.6	0.104	0.002	
22.62	678.50466	n/a	1.49	0.113	0.002	2
11.9	404.20494	1.38	1.11	0.114	0.003	
15.07	304.29863	n/a	1.18	0.109	0.003	
22.72	778.53573	n/a	1.3	0.107	0.003	
5.7	307.1138	n/a	1.52	0.123	0.003	3
5.7	302.15845	1.3	0.99	0.125	0.003	
18.33	489.22543	n/a	1.38	0.124	0.003	
14.65	518.32199	n/a	1.25	0.123	0.004	4
12.84	286.14236	0.99	0.8	0.122	0.004	
14.07	432.20027	1.41	1.41	0.119	0.004	
18.34	534.28301	1.39	1.46	0.142	0.004	
5.7	285.13195	1.33	1.41	0.167	0.005	
12.59	994.39684	n/a	0.88	0.162	0.005	
16.84	480.33336	1.32	1.29	0.173	0.006	
0.89	181.07109	n/a	1.33	0.194	0.007	5

Table A.11: Up-regulated metabolites detected in the TCF7L2 positive mode dataset. All candidates with a PFP of up to 20 % were extracted from the RankProd results. Fold changes are given either as reported by RankProd or computed subsequently on the basis of the group median levels. Grayed rows were excluded from the analysis since they exhibited invalid fold changes (less than 1.0). For the look-up of potential metabolite identifications, the ID index refers to the corresponding row in table A.12.

ID index	centroid rt [min]	centroid m/z [Th]	empirical formula	HMDB ID	Compound name	mass error [ppm]	adduct
1	14.6	476.27548	$C_{21}H_{44}NO_7P$	HMDB11473	LPE(0:0/16:0)	1.63	$[M+Na]^+$
				HMDB11503	LPE(16:0/0:0)		
2	22.62	678.50466	$C_{36}H_{72}NO_8P$	HMDB07866	PC(14:0/14:0)	-2.66	$[M+H]^+$
				HMDB08890	PE(15:0/16:0)		
				HMDB08922	PE(16:0/15:0)		
3	5.7	307.1138	$C_{11}H_{15}NO_8$	HMDB12266	N-succinyl-2-amino-6-ketopimelate	0.70	$[M+NH_4]^+$
4	14.65	518.32199	$C_{24}H_{50}NO_7P$	HMDB10382	LPC(16:0)	0.80	$[M+Na]^+$
5	0.89	181.07109	$C_6H_{12}O_6$	HMDB00122	D-glucose	2.60	$[M+H]^+$
				HMDB00143	D-galactose		
			…	…	…		
			$C_3H_6O_3$	HMDB00190	L-lactic acid	2.60	$[2M+H]^+$
				HMDB01051	Glyceraldehyde		
				…	…		

Table A.12: Putative AMS IDs of up-regulated candidates from positive mode data. The ID indices are linked to table A.11 for reference. In case of ID 5, the list of putative metabolites contains further isobaric hexose sugars ($C_6H_{12}O_6$) or three carbon molecules ($C_3H_6O_3$) that are not shown here.

centroid rt [min]	centroid m/z [Th]	fold change (RankProd)	fold change (medians)	PFP	p-value	ID index
16.55	298.3455	n/a	0.19	0.000	0.000	
14.13	312.15813	n/a	0.44	0.000	0.000	1
12.47	286.14249	n/a	0.54	0.000	0.000	
12.62	450.31956	0.72	0.58	0.003	0.000	2
15.95	548.33272	n/a	0.73	0.003	0.000	
10.67	223.09544	n/a	1.26	0.008	0.000	
10.32	244.18961	0.67	0.50	0.008	0.000	3
10.32	226.17912	0.68	0.58	0.009	0.000	
12.61	432.30849	0.78	0.61	0.022	0.000	
0.82	144.10113	n/a	0.97	0.023	0.000	
8.4	167.99279	n/a	0.77	0.025	0.000	
12.84	286.14236	0.99	0.80	0.031	0.000	
15.47	296.25699	0.72	0.77	0.087	0.001	4
16.08	570.3531	n/a	0.97	0.114	0.002	5
12.63	510.26397	0.86	0.95	0.122	0.002	
12.89	251.12665	0.82	1.19	0.116	0.002	
14.82	287.62431	n/a	0.92	0.185	0.003	
17.9	336.3249	0.78	0.66	0.177	0.003	
13.84	358.29378	n/a	0.92	0.169	0.003	6
14.51	500.27525	n/a	1.07	0.162	0.003	
0.83	182.08027	n/a	0.97	0.168	0.004	7
14.63	343.33057	n/a	0.80	0.174	0.004	8
12.92	450.31954	n/a	0.96	0.170	0.004	
23.2	826.53285	n/a	0.74	0.199	0.005	9
0.79	302.87991	n/a	0.75	0.196	0.005	

Table A.13: Down-regulated metabolites detected in the TCF7L2 positive mode dataset. All candidates with a PFP of up to 20 % were extracted from the RankProd results. Fold changes are given either as reported by RankProd or computed subsequently on the basis of the group median levels. Grayed rows were excluded from the analysis since they exhibited invalid fold changes (greater than 1.0). For the look-up of potential metabolite identifications, the ID index refers to the corresponding row in table A.14.

ID index	centroid rt [min]	centroid m/z [Th]	empirical formula	HMDB ID	Compound name	mass error [ppm]	adduct
1	14.13	312.15813	$C_{14}H_{27}NO_4$	HMDB13238	Heptanoylcarnitine (*)	3.61	$[M+K]^+$
2	12.62	450.31956	$C_{26}H_{43}NO_5$	HMDB00631 HMDB00637 HMDB00708 HMDB06898	Deoxycholic acid glycine conjugate Chenodeoxycholic acid glycine conjugate Glycoursodeoxycholic acid Chenodeoxyglycocholic acid	-4.16	$[M+H]^+$
3	10.32	244.18961	$C_{13}H_{25}NO_3$	HMDB13286	N-undecanoylglycine	-4.40	$[M+H]^+$
4	16.08	570.3531	$C_{28}H_{54}NO_7P$	HMDB10392	LPC(20:2(11Z,14Z))	0.31	$[M+Na]^+$
5	13.84	358.29378	$C_{10}H_{18}O_2$	HMDB04980 HMDB10726	*cis*-4-decenoic acid (*) *trans*-dec-2-enoic acid	-4.07	$[2M+NH_4]^+$
6	0.83	182.08027	$C_9H_{11}NO_3$	HMDB00158 HMDB01119 HMDB02184 HMDB06050	L-tyrosine 4-Hydroxy-4-(3-pyridyl)-butanoic acid (*) L-threo-3-phenylserine o-tyrosine	-4.86	$[M+H]^+$
7	14.63	343.33057	$C_{20}H_{39}NO_2$	HMDB02088	N-oleoylethanolamine	-3.95	$[M+NH_4]^+$

8	23.2	826.53285	$C_{46}H_{78}NO_8P$	HMDB08023
				HMDB08149
				…
					PC(16:1/22:6) (*)
					PC(18:2/20:5) (*)
					…
						-3.20	$[M+Na]^+$

Table A.14: Putative AMS IDs of down-regulated candidates from positive mode data. The ID indices are linked to table A.13 for reference. In the case of ID 8, the list of putative metabolites contains another 14 isobaric PCs ($C_{46}H_{78}NO_8P$) that differ in the double bond configurations of their unsaturated acyl chains (not shown here). (*) These IDs were marked as potential false positives by the rt model. This was true for all the alternatives of ID 8.

centroid rt [min]	centroid m/z [Th]	fold change (RankProd)	fold change (medians)	PFP	p-value	ID index
24.96	297.04649	n/a	1.58	0.001	0.000	
18.89	595.49284	n/a	1.72	0.001	0.000	1
20.43	561.48768	n/a	1.45	0.041	0.000	2
19.64	577.48237	n/a	1.48	0.035	0.000	
4.03	329.12355	1.41	1.54	0.038	0.000	3
21.68	381.37303	1.52	1.76	0.087	0.001	4
19.66	379.23325	n/a	1.73	0.110	0.001	
17.92	474.26111	1.47	1.35	0.097	0.001	5
18.51	593.47656	1.33	1.72	0.096	0.001	
3.05	165.97892	1.15	0.91	0.122	0.002	
0.86	212.00209	n/a	1.24	0.142	0.002	6
18.7	573.45097	n/a	1.65	0.140	0.002	7
18.19	578.30824	1.43	1.21	0.131	0.002	
4.02	283.11816	1.36	1.29	0.134	0.002	
14.97	391.28474	n/a	0.79	0.127	0.002	
9.67	313.1189	n/a	1.06	0.127	0.003	
15.29	391.28471	n/a	1.33	0.128	0.003	8
11.43	265.10781	1.35	1.06	0.131	0.003	
23.52	738.50663	1.41	1.01	0.136	0.003	9
21.97	395.38864	1.45	1.22	0.135	0.003	10
23.59	766.52214	n/a	1.12	0.173	0.005	
4.88	213.02233	n/a	1.25	0.180	0.005	

Table A.15: Up-regulated metabolites detected in the TCF7L2 negative mode dataset. All candidates with a PFP of up to 20 % were extracted from the RankProd results. Fold changes are given either as reported by RankProd or computed subsequently on the basis of the group median levels. Grayed rows were excluded from the analysis since they exhibited invalid fold changes (less than 1.0). For the look-up of potential metabolite identifications, the ID index refers to the corresponding row in table A.16.

ID index	centroid rt [min]	centroid m/z [Th]	empirical formula	HMDB ID	Compound name	mass error [ppm]	adduct
1	18.85	595.49301	$C_{36}H_{68}O_6$	HMDB31089	Glycerol triundecanoate	-2.18	$[M-H]^-$

2	20.41	561.48787	$C_{36}H_{68}O_5$	HMDB07072
				HMDB07184
				...
					DG(15:0/18:1/0:0)
					DG(18:1/15:0/0:0)
					...
						-1.66	$[M-H_2O-H]^-$

3	3.99	329.12361	$C_6H_7N_5O$	HMDB00897 HMDB01566 ...	7-Methylguanine 3-Methylguanine ...	2.33	$[2M-H]^-$
4	21.65	381.37308	$C_{25}H_{50}O_2$	HMDB02361	Pentacosanoic acid	-1.89	$[M-H]^-$

5	17.84	474.26130	$C_{23}H_{42}NO_7P$	HMDB11478
				HMDB11508
				...
					LPE(0:0/18:3) (*)
					LPE(18:3/0:0) (*)
					...
						-2.76	$[M-H]^-$

6	0.86	212.00214	$C_8H_7NO_4S$	HMDB00682	Indoxyl sulfate (*)	-0.77	$[M-H]^-$

7	18.61	573.45110	$C_{36}H_{62}O_5$	HMDB07077
				HMDB07329
				...
					DG(15:0/18:4/0:0) (*)
					DG(18:4/15:0/0:0) (*)
					...
						-2.36	$[M-H]^-$

8	14.93	391.28486	$C_{24}H_{40}O_4$	HMDB00361 HMDB00384 ...	3b,7a-Dihydroxy-5b-cholanoic acid 3a,7a-Dihydroxycholanoic acid ...	-1.33	$[M-H]^-$

9	23.44	738.50674	$C_{41}H_{74}NO_8P$	HMDB08844
				HMDB08937
				...
					PE(14:0/22:4)
					PE(16:0/20:4)
					...
						-1.60	$[M-H]^-$

10	21.92	395.38864	$C_{26}H_{52}O_2$	HMDB02356	Hexacosanoic acid	-2.05	$[M-H]^-$

Table A.16: Putative AMS IDs of up-regulated candidates from the TCF7L2 negative mode data. The ID indices are linked to table A.15 for reference. In the case of IDs 2, 5, 7 and 9, the list of putative metabolites contains other isobaric compounds that differ in the double bond configurations of their unsaturated acyl chains (not shown here). The IDs 3 and 8 refer to other methylguanines and a wide array of bile acid isomers, respectively. (*) These IDs were marked as potential false positives by the rt model. This was true for all the alternatives of IDs 5 and 7.

centroid rt [min]	centroid m/z [Th]	fold change (RankProd)	fold change (medians)	PFP	p-value	ID index
1.92	239.09208	n/a	0.67	0.004	0.000	
8.87	378.10088	n/a	0.48	0.002	0.000	
12.28	399.22015	0.54	0.73	0.007	0.000	
12.75	391.28468	n/a	0.82	0.016	0.000	1
15.05	481.27975	n/a	0.55	0.014	0.000	
25.94	313.0778	n/a	0.91	0.021	0.000	
1.29	172.99132	0.64	0.62	0.023	0.000	
11.06	254.62258	0.64	0.68	0.057	0.001	
11.86	369.17332	0.63	0.82	0.053	0.001	
13.77	613.35797	n/a	0.68	0.050	0.001	
17.46	445.33155	n/a	0.87	0.054	0.001	2
7.7	242.17578	0.69	0.74	0.069	0.001	3
4.88	213.02233	n/a	1.25	0.088	0.002	
10.62	369.17339	0.66	0.78	0.092	0.002	4
5.93	223.13358	n/a	1.01	0.093	0.002	
0.87	172.99118	n/a	0.75	0.113	0.002	
5.94	267.1233	n/a	0.98	0.110	0.002	
14.97	391.28474	n/a	0.79	0.107	0.003	
5.76	443.1549	n/a	0.89	0.165	0.004	
2.16	187.00683	0.84	0.77	0.164	0.004	5
11.27	371.18913	n/a	0.76	0.183	0.005	
14.81	448.30598	0.82	0.54	0.191	0.005	6
14.73	498.28794	n/a	0.84	0.185	0.005	7
0.86	271.06951	0.76	0.85	0.199	0.006	8
15.91	449.2536	0.80	0.83	0.191	0.006	9
10.91	427.21518	0.73	0.89	0.184	0.006	

Table A.17: Down-regulated metabolites detected in the TCF7L2 negative mode dataset. All candidates with a PFP of up to 20 % were extracted from the RankProd results. Fold changes are given either as reported by RankProd or computed subsequently on the basis of the group median levels. Grayed rows were excluded from the analysis since they exhibited invalid fold changes (less than 1.0). For the look-up of potential metabolite identifications, the ID index refers to the corresponding row in table A.18.

ID index	centroid rt [min]	centroid m/z [Th]	empirical formula	HMDB ID	Compound name	mass error [ppm]	adduct
1	12.7	391.28487	$C_{24}H_{40}O_4$	HMDB00518 HMDB00626 ...	CDCA DCA ...	-1.31	$[M-H]^-$
2	17.43	445.33169	$C_{28}H_{48}O_5$	HMDB02163	Trihydroxycoprostanoic acid	-1.38	$[M-H_2O-H]^-$
3	7.61	242.17586	$C_{13}H_{25}NO_3$	HMDB13286	N-undecanoylglycine	-1.26	$[M-H]^-$
4	10.59	369.17356	$C_{19}H_{30}O_5S$	HMDB02759 HMDB06278	Androsterone sulfate 5a-Dihydrotestosterone sulfate	-1.50	$[M-H]^-$
5	2.13	187.00690	$C_7H_8O_4S$	HMDB11635	p-Cresol sulfate	-0.83	$[M-H]^-$
6	14.75	448.30600	$C_{26}H_{43}NO_5$	HMDB00637 HMDB00631	CDCA glycine conjugate DCA glycine conjugate	-1.88	$[M-H]^-$
7	14.66	498.28823	$C_{26}H_{45}NO_6S$	HMDB00896 HMDB00951	TDCA TCDCA	-2.51	$[M-H]^-$
8	0.86	271.06960	$C_5H_4N_4O$	HMDB00157	Hypoxanthine	-0.53	$[2M-H]^-$
9	15.86	449.25373	$C_{25}H_{40}O_8$	HMDB10321 HMDB10339 HMDB10359	3,17-Androstanediol gluc. 3-alpha-androstanediol gluc. (*) 17-hydroxyandrostane-3-gluc. (*)	-1.59	$[M-H_2O-H]^-$

Table A.18: Putative AMS IDs of down-regulated candidates from the TCF7L2 negative mode data. The ID indices are linked to table A.17 for reference. In the case of ID 1, the list of putative metabolites contains other isobaric bile acids (not shown here). (*) These IDs were marked as potential false positives by the rt model.

APPENDIX B

Curriculum Vitae

Education

07/2008 – 04/2014 **PhD candidate**	Applied Bioinformatics (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls University of Tübingen**, Germany
06/2007 – 12/2007 **Diploma thesis**	Simulation of Biological Systems (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls University of Tübingen**, Germany
10/2000 – 04/2008 **Bioinformatics studies**	**Eberhard Karls University of Tübingen**, Germany

Scientific Work Experience

11/2012 **Research visit**	Division of Chemistry and Bioanalytics (Prof. Christian G. Huber), **University of Salzburg**, Austria
08/2012 – 12/2013 **System administration**	Applied Bioinformatics (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls University of Tübingen**, Germany

12/2010 **Research visit**	Sino German Cooperation, CAS Key Laboratory of Separation Science for Analytical Chemistry (Prof. Guowang Xu), **Dalian Institute of Chemical Physics**, China
2008 – 2011 **Regular research visits**	Research Unit Analytic Biogeochemistry, **Helmholtz-Zentrum München**, Germany
07/2008 – 07/2013 **Teaching assistant**	Tutoring of coursework accompanying lectures *Computational Drug Design I & II* and software engineering practical courses, Applied Bioinformatics (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls University of Tübingen**, Germany
07/2008 – 07/2013 **Software development** (OpenMS developer)	Applied Bioinformatics (Prof. Oliver Kohlbacher), Center for Bioinformatics, **Eberhard Karls University of Tübingen**, Germany

Publications

- S. Aiche, T. Sachsenberg, E. Kenar, M. Walzer, B. Wiswedel, T. Kristl, M. Boyles, A. Duschl, C. G. Huber, M. R. Berthold, K. Reinert, and O. Kohlbacher. "Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry". In: *PROTEOMICS* (2015), epub ahead of print. DOI: 10.1002/pmic.201400391.
- R. Wagner, J. Li, E. Kenar, O. Kohlbacher, F. Machicao, H.-U. Häring, A. Fritsche, G. Xu, and R. Lehmann. "Clinical and non-targeted metabolomic profiling of homozygous carriers of Transcription Factor 7-like 2 variant rs7903146". In: *Scientific Reports* 4 (2014), article number 5296. DOI: 10.1038/srep05296.
- E. Kenar, H. Franken, S. Forcisi, K. Wörmann, H.-U. Häring, R. Lehmann, P. Schmitt-Kopplin, A. Zell, and O. Kohlbacher. "Automated Label-free Quantification of Metabolites from Liquid Chromatography–Mass Spectrometry Data". In: *Molecular & Cellular Proteomics* 13 (2014), pp. 348–359. DOI: 10.1074/mcp.M113.031278.
- H. Weisser, S. Nahnsen, J. Grossmann, L. Nilse, A. Quandt, H. Brauer, M. Sturm, E. Kenar, O. Kohlbacher, R. Aebersold, and L. Malmström. "An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics". In: *Journal of Proteome Research* 12.4 (2013), pp. 1628–1644. DOI: 10.1021/pr300992u.
- E. Kenar, H. Franken, L. Rosenbaum, R. Lehmann, S. Forcisi, K. Wörmann, M. Lucio, A. König, J. Rahnenführer, P. Schmitt-Kopplin, H.-U. Häring, A. Zell, and O. Kohlbacher. "Mit Bioinformatik zu Biomarkern". In: *Medizinische Welt* 63.5 (2012), pp. 245–250.

- K. Wörmann, M. Lucio, S. Forcisi, S. S. Heinzmann, E. Kenar, H. Franken, L. Rosenbaum, P. Schmitt-Kopplin, O. Kohlbacher, A. Zell, H.-U. Häring, and R. Lehmann. "„Metabolomics“ in der Diabetesforschung". In: *Der Diabetologe* 8.1 (2012), pp. 42–48. DOI: 10.1007/s11428-011-0778-9.

Manuscripts in Preparation

- E. Kenar, J. Hansen, X. Zhao, S. Chen, X. Li, H.-U. Häring, G. Xu, B. K. Pedersen, R. Lehmann, O. Kohlbacher, P. Plomgaard, and C. Weigert. "A combined non-targeted metabolomics bioinformatics strategy reveals the dynamics of metabolic patterns and affected metabolic pathways induced by one single bout of endurance exercise in healthy men".

Posters

- E. Kenar, H. Franken, A. Fekete, S. Forcisi, K. Wörmann, Ph. Schmitt-Kopplin, R. Lehmann, H.-U. Häring, A. Zell, and O. Kohlbacher. "Metabolomic Feature Detection in LC-MS data — A Probabilistic Approach". *Metabolomics & More Symposium* (2010), Freising-Weihenstephan, Germany.

Diploma Thesis

- E. Kenar. "Modellierung von Wechselwirkungen zwischen DNA-Aptameren und Thrombin". *Diploma thesis* (2007). Simulation of Biological Systems (Prof. Oliver Kohlbacher), Center for Bioinformatics, Eberhard Karls University of Tübingen, Germany.

References

[1] *Comparison of a single-stranded RNA and a double-stranded DNA with their corresponding nucleobases*. 2010. URL: http://commons.wikimedia.org/wiki/File:Difference%5C_DNA%5C_RNA-DE.svg.

[2] W. Bode et al. "The refined 1.9 A crystal structure of human alpha-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment." In: *The EMBO journal* 8.11 (Nov. 1989), pp. 3467–3475. DOI: 10.2210/pdb1ppb/pdb.

[3] *The Human Metabolome Database*. 2014. URL: http://www.hmdb.ca.

[4] G. Michal. *Biochemical Pathways: an atlas of biochemistry and molecular biology*. New York: Wiley, 1999. ISBN: 0-471-33130-9.

[5] B. Kuska. "Beer, Bethesda, and biology: how "genomics" came into being." In: *Journal of the National Cancer Institute* 90.2 (1998), p. 93. DOI: 10.1093/jnci/90.2.93.

[6] K. Dettmer, P. a. Aronov, and B. D. Hammock. "Mass spectrometry-based metabolomics". In: *Mass Spectrometry Reviews* 26.1 (2007), pp. 51–78. DOI: 10.1002/mas.20108.

[7] M. Monteiro et al. "Metabolomics Analysis for Biomarker Discovery: Advances and Challenges". In: *Current Medicinal Chemistry* 20.2 (2012), pp. 257–271. DOI: 10.2174/0929867311320020006.

[8] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2008. ISBN: 3-900051-07-0.

[9] C. A. Smith et al. "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification". In: *Analytical Chemistry* 78.3 (Feb. 2006), pp. 779–787. ISSN: 1520-6882. DOI: 10.1021/ac051437y. URL: http://dx.doi.org/10.1021/ac051437y.

[10] C. Kuhl et al. “CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets”. In: *Analytical Chemistry* 84.1 (2012), pp. 283–289. DOI: 10.1021/ac202450g.

[11] M. Sturm et al. “OpenMS - an open-source software framework for mass spectrometry.” In: *BMC bioinformatics* 9.1 (2008), p. 163. DOI: 10.1186/1471-2105-9-163.

[12] L. W. Sumner et al. “Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)”. In: *Metabolomics* 3.3 (2007), pp. 211–221. DOI: 10.1007/s11306-007-0082-2.

[13] H. Horai et al. “MassBank: A public repository for sharing mass spectral data for life sciences”. In: *Journal of Mass Spectrometry* 45.7 (2010), pp. 703–714. DOI: 10.1002/jms.1777.

[14] J. Griss et al. “The mzTab Data Exchange Format: communicating MS-based proteomics and metabolomics experimental results to a wider audience.” In: *Molecular & cellular proteomics : MCP* (2014), pp. 1–28. DOI: 10.1074/mcp.O113.036681.

[15] S. Bijlsma et al. “Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation”. In: *Analytical Chemistry* 78.2 (2006), pp. 567–574. DOI: 10.1021/ac051495j.

[16] Y. Benjamini and Y. Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing.” In: *Journal of the Royal Statisitical Society, Series B* 57.1 (1995), pp. 289–300. URL: http://links.jstor.org/sici?sici=0035-9246(1995)57:1%3C289:CTFDRA%3E2.0.CO;2-E%5C&origin=MSN.

[17] E. Szymańska et al. “Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies”. In: *Metabolomics* 8.S1 (2012), pp. 3–16. DOI: 10.1007/s11306-011-0330-3.

[18] R. Breitling et al. “Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments”. In: *FEBS Letters* 573.1-3 (Aug. 2004), pp. 83–92. DOI: 10.1016/j.febslet.2004.07.055.

[19] R. Wagner et al. “Clinical and non-targeted metabolomic profiling of homozygous carriers of Transcription Factor 7-like 2 variant rs7903146.” In: *Scientific reports* 4 (2014), p. 5296. DOI: 10.1038/srep05296.

[20] J. Prawitt, S. Caron, and B. Staels. “Bile acid metabolism and the pathogenesis of type 2 diabetes”. In: *Current Diabetes Reports* 11.3 (2011), pp. 160–166. DOI: 10.1007/s11892-011-0187-x.

[21] G. Corona et al. “Sexual dysfunction at the onset of type 2 diabetes: The interplay of depression, hormonal and cardiovascular factors”. In: *Journal of Sexual Medicine* 11.8 (Aug. 2014), pp. 2065–2073. DOI: 10.1111/jsm.12601.

[22] G. Corona et al. "Following the common association between testosterone deficiency and diabetes mellitus, can testosterone be regarded as a new therapy for diabetes?" In: *International Journal of Andrology* 32.5 (2009), pp. 431–441. DOI: 10.1111/j.1365-2605.2009.00965.x.

[23] T. Scharl and F. Leisch. "The Stochastic QT–clust Algorithm: Evaluation of Stability and Variance on Time–course Microarray Data". In: *Compstat 2006—Proceedings in Computational Statistics*. Ed. by A. Rizzi and M. Vichi. Proceedings in Computational Statistics. Physica Verlag, Heidelberg, Germany, 2006, pp. 1015–1022. ISBN: 3-7908-1708-2.

[24] C. Möller-Levet et al. "Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points". English. In: *Advances in Intelligent Data Analysis V*. Ed. by M. R. Berthold et al. Vol. 2810. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, pp. 330–340. ISBN: 978-3-540-40813-0. DOI: 10.1007/978-3-540-45231-7_31.

[25] R. J. Bino et al. "Potential of metabolomics as a functional genomics tool". In: *Trends in Plant Science* 9.9 (2004), pp. 418–425. DOI: 10.1016/j.tplants.2004.07.004.

[26] D. S. Wishart et al. "HMDB: A knowledgebase for the human metabolome". In: *Nucleic Acids Research* 37.suppl 1 (2009), pp. D603–610. DOI: 10.1093/nar/gkn810.

[27] D. S. Wishart et al. "HMDB 3.0-The Human Metabolome Database in 2013". In: *Nucleic Acids Research* 41.Database issue (2013), pp. D801–7. DOI: 10.1093/nar/gks1065.

[28] W. Weckwerth. "Metabolomics in systems biology." In: *Annual review of plant biology* 54.1 (2003), pp. 669–689. DOI: 10.1146/annurev.arplant.54.031902.135014.

[29] J. K. Nicholson and I. D. Wilson. "Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism." In: *Nature reviews. Drug discovery* 2.8 (2003), pp. 668–676. DOI: 10.1038/nrd1157.

[30] E. M. Lenz and I. D. Wilson. "Analytical strategies in metabonomics". In: *Journal of Proteome Research* 6.2 (2007), pp. 443–458. DOI: 10.1021/pr0605217.

[31] J. M. Halket and V. G. Zaikin. "Derivatization in mass spectrometry–1. Silylation." eng. In: *Eur J Mass Spectrom (Chichester, Eng)* 9.1 (2003), pp. 1–21. DOI: 10.1255/ejms.527.

[32] R. Breitling, A. R. Pitt, and M. P. Barrett. "Precision mapping of the metabolome". In: *Trends in Biotechnology* 24.12 (2006), pp. 543–548. DOI: 10.1016/j.tibtech.2006.10.006.

[33] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson. "Fourier transform ion cyclotron resonance mass spectrometry: a primer." In: *Mass spectrometry reviews* 17.1 (1998), pp. 1–35. DOI: 10.1002/(SICI)1098-2787(1998)17:1<1::AID-MAS1>3.0.CO;2-K.

[34] Q. Hu et al. "The Orbitrap: A new mass spectrometer". In: *Journal of Mass Spectrometry* 40.4 (2005), pp. 430–443. DOI: 10.1002/jms.856.

[35] A. Furey et al. "Ion suppression; A critical review on causes, evaluation, prevention and applications". In: *Talanta* 115 (2013), pp. 104–122. DOI: 10.1016/j.talanta.2013.03.048.

[36] Snyder, Lloyd R. and Kirkland, J. J. and Dolan, John W. *Introduction to modern liquid chromatography*. 2010, p. 912. ISBN: 0471038229. DOI: 10.1002/9780470508183.

[37] J. B. Fenn et al. "Electrospray ionization for mass spectrometry of large biomolecules." In: *Science (New York, N.Y.)* 246.4926 (1989), pp. 64–71. DOI: 10.1126/science.2675315.

[38] Lord Rayleigh. "XX. On the equilibrium of liquid conducting masses charged with electricity". In: *Philosophical Magazine Series 5* 14.87 (Sept. 1882), pp. 184–186. DOI: 10.1080/14786448208628425.

[39] G. Siuzdak. "Mass Analyzers and Ion Detectors". In: *Mass Spectrometry for Biotechnology*. Elsevier BV, 1996, pp. 32–55. DOI: 10.1016/b978-012647471-8/50005-3.

[40] K. H. Kingdon. "A method for the neutralization of electron space charge by positive ionization at very low gas pressures". In: *Physical Review* 21.4 (Apr. 1923), pp. 408–418. DOI: 10.1103/PhysRev.21.408.

[41] R. D. Knight. "Storage of ions from laser-produced plasmas". In: *Applied Physics Letters* 38.4 (1981), pp. 221–223. DOI: 10.1063/1.92315.

[42] A. Makarov. "Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis". In: *Analytical Chemistry* 72.6 (2000), pp. 1156–1162. DOI: 10.1021/ac991131p.

[43] A. Michalski et al. "Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer." In: *Molecular & cellular proteomics : MCP* 10.9 (2011), p. M111.011015. DOI: 10.1074/mcp.M111.011015.

[44] J. V. Olsen et al. "Higher-energy C-trap dissociation for peptide modification analysis." In: *Nature methods* 4.9 (Sept. 2007), pp. 709–712. DOI: 10.1038/nmeth1060.

[45] Thermo Fisher Scientific. *Exactive(tm) Operating Manual*. 2010. URL: http://www.thermoscientific.com/content/dam/tfs/ATG/CMD/cmd-support/q-exactive/operations-and-maintenance/operators-manuals/1249360-Exactive-Operating-Rev-C.pdf.

[46] O. Fiehn. "Metabolomics - The link between genotypes and phenotypes". In: *Plant Molecular Biology* 48.1-2 (2002), pp. 155–171. DOI: 10.1023/A:1013713905833.

[47] K. Rangiah et al. "Nicotine exposure and metabolizer phenotypes from analysis of urinary nicotine and its 15 metabolites by LC-MS." In: *Bioanalysis* 3.7 (Apr. 2011), pp. 745–761. DOI: 10.4155/BIO.11.42.

[48] P. Giavalisco et al. “13C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research”. In: *Analytical Chemistry* 81.15 (Aug. 2009), pp. 6546–6551. DOI: 10.1021/ac900979e.

[49] M. R. Shortreed et al. “Ionizable isotopic labeling reagent for relative quantification of amine metabolites by mass spectrometry”. In: *Analytical Chemistry* 78.18 (Sept. 2006), pp. 6398–6403. DOI: 10.1021/ac0607008.

[50] S. Nahnsen et al. “Tools for Label-free Peptide Quantification”. In: *Molecular & Cellular Proteomics* 12.3 (Dec. 2012), pp. 549–556. DOI: 10.1074/mcp.r112.025163.

[51] W. Zhu, J. W. Smith, and C. M. Huang. “Mass spectrometry-based label-free quantitative proteomics”. In: *Journal of Biomedicine and Biotechnology* 2010 (2010), pp. 1–6. DOI: 10.1155/2010/840518.

[52] J. P. Foley. “Equations for chromatographic peak modeling and calculation of peak area”. In: *Analytical Chemistry* 59.15 (1987), pp. 1984–1987. DOI: 10.1021/ac00142a019.

[53] E. Kenar et al. “Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data.” In: *Molecular & cellular proteomics : MCP* 13.1 (2014), pp. 348–59. DOI: 10.1074/mcp.M113.031278.

[54] E. Lange et al. “Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements”. In: *BMC Bioinformatics* 9.1 (2008), p. 375. DOI: 10.1186/1471-2105-9-375.

[55] H. Weisser et al. “An automated pipeline for high-throughput label-free quantitative proteomics”. In: *Journal of Proteome Research* 12.4 (2013), pp. 1628–1644. DOI: 10.1021/pr300992u. URL: http://www.ncbi.nlm.nih.gov/pubmed/23391308.

[56] O. Kohlbacher et al. “TOPP - The OpenMS proteomics pipeline”. In: *Bioinformatics* 23.2 (2007), e191–e197. DOI: 10.1093/bioinformatics/btl299.

[57] L. Martens et al. “mzML–a community standard for mass spectrometry data.” In: *Molecular & cellular proteomics : MCP* 10.1 (2011), R110.000133. DOI: 10.1074/mcp.R110.000133.

[58] D. Kessner et al. “ProteoWizard: Open source software for rapid proteomics tools development”. In: *Bioinformatics* 24.21 (2008), pp. 2534–2536. DOI: 10.1093/bioinformatics/btn323.

[59] *Biocrates Life Sciences AG*. 2014. URL: http://www.biocrates.com.

[60] E. Altmaier et al. “Bioinformatics analysis of targeted metabolomics - Uncovering old and new tales of diabetic mice under medication”. In: *Endocrinology* 149.7 (2008), pp. 3478–3489. DOI: 10.1210/en.2007-1747.

[61] D. S. Wishart. “Advances in metabolite identification.” In: *Bioanalysis* 3.15 (Aug. 2011), pp. 1769–1782. DOI: 10.4155/bio.11.155.

[62] G. S. Lueker. *Two NP-complete Problems in Nonnegative Integer Programming*. Technical Report TR-178. Department of Electrical Engineering, Princeton University, 1975.

[63] S. Martello and P. Toth. *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc., 1990. ISBN: 978-0471924203.

[64] S. Böcker and Z. Lipták. "Efficient mass decomposition". In: *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*. ACM. 2005, pp. 151–157. ISBN: 1581139640. DOI: 10.1145/1066677.1066715.

[65] T. Kind and O. Fiehn. "Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm." In: *BMC Bioinformatics* 7.1 (2006), p. 234. DOI: 10.1186/1471-2105-7-234.

[66] T. Kind and O. Fiehn. "Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry." In: *BMC Bioinformatics* 8 (2007), p. 105. DOI: 10.1186/1471-2105-8-105.

[67] M. Brown et al. "Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics." In: *The Analyst* 134.7 (2009), pp. 1322–1332. DOI: 10.1039/b901179j.

[68] J. Draper et al. "Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'." In: *BMC Bioinformatics* 10.1 (2009), p. 227. DOI: 10.1186/1471-2105-10-227.

[69] G. T. Gipson et al. "Assignment of MS-based metabolomic datasets via compound interaction pair mapping". In: *Metabolomics* 4.1 (2008), pp. 94–103. DOI: 10.1007/s11306-007-0096-9.

[70] S. Rogers et al. "Probabilistic assignment of formulas to mass peaks in metabolomics experiments". In: *Bioinformatics* 25.4 (2009), pp. 512–518. DOI: 10.1093/bioinformatics/btn642.

[71] R. J. M. Weber and M. R. Viant. "MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways". In: *Chemometrics and Intelligent Laboratory Systems* 104.1 (2010), pp. 75–82. DOI: 10.1016/j.chemolab.2010.04.010.

[72] W. B. Dunn et al. "Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry." In: *Nature protocols* 6.7 (2011), pp. 1060–1083. DOI: 10.1038/nprot.2011.335.

[73] T. Kind et al. "FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry". In: *Analytical Chemistry* 81.24 (2009), pp. 10038–10048. DOI: 10.1021/ac9019522.

[74] J. Lisec et al. "Gas chromatography mass spectrometry-based metabolite profiling in plants." In: *Nature protocols* 1.1 (2006), pp. 387–396. DOI: 10.1038/nprot.2006.59.

[75] J. Kopka et al. “GMD@CSB.DB: The Golm metabolome database”. In: *Bioinformatics* 21.8 (2005), pp. 1635–1638. DOI: 10.1093/bioinformatics/bti236.

[76] R. Kaliszan et al. “Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships”. In: *Proteomics* 5.2 (2005), pp. 409–415. DOI: 10.1002/pmic.200400973.

[77] N. Pfeifer et al. “Improving peptide identification in proteome analysis by a two-dimensional retention time filtering approach”. In: *Journal of Proteome Research* 8.8 (2009), pp. 4109–4115. DOI: 10.1021/pr900064b.

[78] D. J. Creek et al. “Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: Improved metabolite identification by retention time prediction”. In: *Analytical Chemistry* 83.22 (2011), pp. 8703–8710. DOI: 10.1021/ac2021823.

[79] M. Sugimoto et al. “Large-scale prediction of cationic metabolite identity and migration time in capillary electrophoresis mass spectrometry using artificial neural networks”. In: *Analytical Chemistry* 77.1 (2005), pp. 78–84. DOI: 10.1021/ac048950g.

[80] L. Sleno and D. a. Volmer. “Ion activation methods for tandem mass spectrometry”. In: *Journal of Mass Spectrometry* 39.10 (2004), pp. 1091–1112. DOI: 10.1002/jms.703.

[81] S. E. Stein and D. R. Scott. “Optimization and testing of mass spectral library search algorithms for compound identification”. In: *Journal of the American Society for Mass Spectrometry* 5.9 (Sept. 1994), pp. 859–866. DOI: 10.1016/1044-0305(94)87009-8.

[82] D. Fenyö and R. C. Beavis. “A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes”. In: *Analytical Chemistry* 75.4 (2003), pp. 768–774. DOI: 10.1021/ac0258709.

[83] R. Craig and R. C. Beavis. “TANDEM: Matching proteins with tandem mass spectra”. In: *Bioinformatics* 20.9 (2004), pp. 1466–1467. DOI: 10.1093/bioinformatics/bth092.

[84] R. Mylonas et al. “X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry.” In: *Analytical chemistry* 81.18 (Sept. 2009), pp. 7604–7610. DOI: 10.1021/ac900954d.

[85] C. a. Smith et al. “METLIN: a metabolite mass spectral database.” In: *Therapeutic drug monitoring* 27.6 (2005), pp. 747–751. DOI: 10.1097/01.ftd.0000179845.53213.39.

[86] M. Heinonen et al. “FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data”. In: *Rapid Communications in Mass Spectrometry* 22.19 (2008), pp. 3043–3052. DOI: 10.1002/rcm.3701.

[87] D. W. Hill et al. “Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra”. In: *Analytical Chemistry* 80.14 (2008), pp. 5574–5582. DOI: 10.1021/ac800548g.

[88] J. Bartel, J. Krumsiek, and F. J. Theis. “Statistical methods for the analysis of high-throughput metabolomics data.” In: *Computational and structural biotechnology journal* 4 (2013), e201301009. DOI: 10.5936/csbj.201301009.

[89] K. Kim et al. “Urine metabolomics analysis for kidney cancer detection and biomarker discovery.” In: *Molecular & cellular proteomics : MCP* 8.3 (2009), pp. 558–570. DOI: 10.1074/mcp.M800165-MCP200.

[90] T. Kind et al. “A comprehensive urinary metabolomic approach for identifying kidney cancer”. In: *Analytical Biochemistry* 363.2 (2007), pp. 185–195. DOI: 10.1016/j.ab.2007.01.028.

[91] J. Trygg and S. Wold. “Orthogonal projections to latent structures (O-PLS)”. In: *Journal of Chemometrics* 16.3 (2002), pp. 119–128. DOI: 10.1002/cem.695.

[92] V. W. Davis et al. “Urinary metabolomic signature of esophageal cancer and Barrett’s esophagus.” In: *World journal of surgical oncology* 10 (2012), p. 271. DOI: 10.1186/1477-7819-10-271.

[93] H. S. Tapp and E. K. Kemsley. “Notes on the practical utility of OPLS”. In: *TrAC - Trends in Analytical Chemistry* 28.11 (2009), pp. 1322–1327. DOI: 10.1016/j.trac.2009.08.006.

[94] J. L. DeRisi, V. R. Iyer, and P. O. Brown. “Exploring the metabolic and genetic control of gene expression on a genomic scale.” In: *Science (New York, N.Y.)* 278.5338 (1997), pp. 680–686. DOI: 10.1126/science.278.5338.680.

[95] V. G. Tusher, R. Tibshirani, and G. Chu. “Significance analysis of microarrays applied to the ionizing radiation response.” In: *Proceedings of the National Academy of Sciences of the United States of America* 98.9 (Apr. 2001), pp. 5116–5121. DOI: 10.1073/pnas.091062498.

[96] H. Sun et al. “Metabolomic Analysis of Key Regulatory Metabolites in Hepatitis C Virus–infected Tree Shrews”. In: *Molecular & Cellular Proteomics* 12.3 (2013), pp. 710–719. DOI: 10.1074/mcp.M112.019141.

[97] G. Smyth. “limma: Linear Models for Microarray Data”. English. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Ed. by R. Gentleman et al. Statistics for Biology and Health. Springer New York, 2005, pp. 397–420. ISBN: 978-0-387-25146-2. DOI: 10.1007/0-387-29362-0_23. URL: http://dx.doi.org/10.1007/0-387-29362-0_23.

[98] J. Alvarez et al. “Plasma metabolomic profiling in adults with cystic fibrosis and cystic fibrosis-related diabetes”. In: *FASEB Journal* 28.1 Supplement (2014). URL: http://www.fasebj.org/content/28/1_Supplement/248.1.

[99] B. M. Bolstad et al. “A comparison of normalization metholds for high density oligonucleotide array data based on variance and bias”. In: *Bioinformatics* 19.2 (2003), pp. 185–193.

[100] B. M. Bolstad. *preprocessCore: A collection of pre-processing functions*. R package version 1.24.0. URL: http://www.bioconductor.org/packages/release/bioc/html/preprocessCore.html.

[101] J. Lee et al. “Quantile Normalization Approach for Liquid Chromatography–Mass Spectrometry-based Metabolomic Data from Healthy Human Volunteers”. In: *Analytical Sciences* 28.8 (2012), pp. 801–805. DOI: 10.2116/analsci.28.801.

[102] F. Hong et al. “RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis”. In: *Bioinformatics* 22.22 (Nov. 2006), pp. 2825–2827. DOI: 10.1093/bioinformatics/btl476.

[103] T. Yu et al. “apLCMS-adaptive processing of high-resolution LC/MS data”. In: *Bioinformatics* 25.15 (2009), pp. 1930–1936. DOI: 10.1093/bioinformatics/btp291.

[104] K. Reinert and O. Kohlbacher. “OpenMS and TOPP: Open Source Software for LC-MS Data Analysis”. In: *Proteome Bioinformatics* (Dec. 2009), pp. 201–211. ISSN: 1940-6029. DOI: 10.1007/978-1-60761-444-9_14.

[105] E. Melamud, L. Vastag, and J. D. Rabinowitz. “Metabolomic analysis and visualization engine for LC - MS data”. In: *Analytical Chemistry* 82.23 (2010), pp. 9818–9826. DOI: 10.1021/ac1021166.

[106] C. Bueschl et al. “Metextract: A new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research”. In: *Bioinformatics* 28.5 (2012), pp. 736–738. DOI: 10.1093/bioinformatics/bts012.

[107] M. Katajamaa and M. Orešič. “Data processing for mass spectrometry-based metabolomics”. In: *Journal of Chromatography A* 1158.1-2 (2007), pp. 318–328. DOI: 10.1016/j.chroma.2007.04.021.

[108] R. Tautenhahn, C. Böttcher, and S. Neumann. “Annotation of LC/ESI-MS Mass Signals”. In: *Lecture Notes in Computer Science* (2007), pp. 371–380. DOI: 10.1007/978-3-540-71233-6_29.

[109] K. M. Åberg et al. “Feature detection and alignment of hyphenated chromatographic-mass spectrometric data. Extraction of pure ion chromatograms using Kalman tracking”. In: *Journal of Chromatography A* 1192.1 (2008), pp. 139–146. DOI: 10.1016/j.chroma.2008.03.033.

[110] M. Katajamaa and M. Orešič. “Processing methods for differential analysis of LC/MS profile data.” In: *BMC bioinformatics* 6 (2005), p. 179. DOI: 10.1186/1471-2105-6-179.

[111] V. P. Andreev et al. “A Universal Denoising and Peak Picking Algorithm for LC-MS Based on Matched Filtration in the Chromatographic Time Domain”. In: *Analytical Chemistry* 75.22 (2003), pp. 6314–6326. DOI: 10.1021/ac0301806.

[112] K. C. Leptos et al. “MapQuant: Open-source software for large-scale protein quantification”. In: *Proteomics* 6.6 (2006), pp. 1770–1782. DOI: 10.1002/pmic.200500201.

[113] T. Pluskal et al. “MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.” In: *BMC Bioinformatics* 11.1 (2010), p. 395. DOI: 10.1186/1471-2105-11-395.

[114] M. Bellew et al. “A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS”. In: *Bioinformatics* 22.15 (2006), pp. 1902–1909. DOI: 10.1093/bioinformatics/btl276.

[115] M. Hermansson et al. “Automated quantitative analysis of complex lipidomes by liquid chromatography/mass spectrometry”. In: *Analytical Chemistry* 77.7 (2005), pp. 2166–2175. DOI: 10.1021/ac048489s.

[116] J. Cox and M. Mann. “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.” In: *Nature biotechnology* 26.12 (2008), pp. 1367–1372. DOI: 10.1038/nbt.1511.

[117] M. W. Senko, S. C. Beu, and F. W. McLaffertycor. “Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions”. In: *Journal of the American Society for Mass Spectrometry* 6.4 (1995), pp. 229–233. DOI: 10.1016/1044-0305(95)00017-8.

[118] J. Junker et al. “TOPPAS: A graphical workflow editor for the analysis of high-throughput proteomics data”. In: *Journal of Proteome Research* 11.7 (2012), pp. 3914–3920. DOI: 10.1021/pr300187f.

[119] V. a. Petyuk et al. “Elimination of systematic mass measurement errors in liquid chromatography-mass spectrometry based proteomics using regression models and a priori partial knowledge of the sample content”. In: *Analytical Chemistry* 80.3 (2008), pp. 693–706. DOI: 10.1021/ac701863d.

[120] S. Dasgupta and D. Hsu. “On-line Estimation with the Multivariate Gaussian Distribution”. In: *Proceedings of the 20th Annual Conference on Learning Theory*. Springer-Verlag, 2007, pp. 278–292. ISBN: 978-3-540-72925-9. URL: http://dl.acm.org/citation.cfm?id=1768841.1768869.

[121] W. S. Cleveland, S. J. Devlin, and S. Cleveland. “Locally Weighted Regression : An Approach to Regression Analysis by Local Fitting”. In: *Journal of the American Statistical Association* 83.403 (2013), pp. 596–610. DOI: 10.2307/2289282.

[122] A. L. Rockwood and P. Haimi. "Efficient calculation of accurate masses of isotopic peaks". In: *Journal of the American Society for Mass Spectrometry* 17.3 (2006), pp. 415–419. DOI: 10.1016/j.jasms.2005.12.001.

[123] B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." In: *Biochimica et biophysica acta* 405.2 (1975), pp. 442–451. DOI: 10.1016/0005-2795(75)90109-9.

[124] The MathWorks Inc. *MATLAB R2012b*. 2012.

[125] P. Giavalisco et al. "Elemental formula annotation of polar and lipophilic metabolites using 13C, 15N and 34S isotope labelling, in combination with high-resolution mass spectrometry". In: *Plant Journal* 68.2 (2011), pp. 364–376. DOI: 10.1111/j.1365-313X.2011.04682.x.

[126] C. Bielow et al. "MSSimulator: Simulation of mass spectrometry data". In: *Journal of Proteome Research* 10.7 (2011), pp. 2922–2929. DOI: 10.1021/pr200155f.

[127] D. C. Blair. *Information Retrieval*. 2nd. Vol. 30. Journal of the American Society for Information Science. London: Wiley Subscription Services, Inc., A Wiley Company, 1979, pp. 374–375. DOI: 10.1002/asi.4630300621.

[128] D. S. Wishart et al. "HMDB: The human metabolome database". In: *Nucleic Acids Research* 35.Database issue (2007), pp. D521–6. DOI: 10.1093/nar/gkl923.

[129] M. R. Berthold et al. "KNIME: the Konstanz Information Miner". In: *SIGKDD Explor. Newsl.* 11.1 (2006), pp. 26–31. DOI: 10.1145/1656274.1656280.

[130] R. J. M. Weber et al. "Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification". In: *Analytical Chemistry* 83.10 (2011), pp. 3737–3743. DOI: 10.1021/ac2001803.

[131] Metabolomics Fiehn Lab. *Mass Spectrometry Adduct Calculator*. 2010. URL: http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/MS-Adduct-Calculator/.

[132] MassBank. *MassBank SVN repository*. 2013. URL: http://www.massbank.jp/SVN/OpenData/record/.

[133] H. L. Röst et al. "pyOpenMS: A Python-based interface to the OpenMS mass-spectrometry algorithm library". In: *Proteomics* 14.1 (2014), pp. 74–77. DOI: 10.1002/pmic.201300246.

[134] *OpenMS 1.11.1 Documentation*. 2013. URL: http://www.openms.de.

[135] K. Biemann. "Mass spectrometry of peptides and proteins." In: *Annual review of biochemistry* 61.1 (1992), pp. 977–1010. DOI: 10.1146/annurev.biochem.61.1.977.

[136] A. Leo, C. Hansch, and D. Elkins. "Partition coefficients and their Uses". In: *Chemical Reviews* 71.6 (1971), pp. 525–616. DOI: 10.1021/cr60274a001.

[137] R. Wang, Y. Gao, and L. Lai. “Calculating partition coefficient by atom-additive method”. In: *Perspectives in Drug Discovery and Design* 19.1 (2000), pp. 47–66. DOI: 10.1023/A:1008763405023.

[138] C. W. Yap. “PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints”. In: *Journal of Computational Chemistry* 32.7 (2011), pp. 1466–1474. DOI: 10.1002/jcc.21707.

[139] *RDKit: Open-source cheminformatics*. URL: http://www.rdkit.org.

[140] S. Beisken et al. “KNIME-CDK: Workflow-driven cheminformatics.” In: *BMC Bioinformatics* 14.1 (2013), p. 257. DOI: 10.1186/1471-2105-14-257.

[141] M. Gerlich and S. Neumann. “MetFusion: Integration of compound identification strategies”. In: *Journal of Mass Spectrometry* 48.3 (2013), pp. 291–298. DOI: 10.1002/jms.3123.

[142] R. Tautenhahn et al. “An accelerated workflow for untargeted metabolomics using the METLIN database”. In: *Nature Biotechnology* 30.9 (Sept. 2012), pp. 826–828. DOI: 10.1038/nbt.2348.

[143] J. K. Kim et al. “Time-course metabolic profiling in Arabidopsis thaliana cell cultures after salt stress treatment”. In: *Journal of Experimental Botany* 58.3 (2007), pp. 415–424. DOI: 10.1093/jxb/erl216.

[144] M. J. Brauer et al. “Conservation of the metabolomic response to starvation across two divergent microbes.” In: *Proceedings of the National Academy of Sciences of the United States of America* 103.51 (2006), pp. 19302–19307. DOI: 10.1073/pnas.0609508103.

[145] E. Chorell et al. “Predictive Metabolomics Evaluation of Nutrition-Modulated Metabolic Stress Responses in Human Blood Serum During the Early Recovery Phase of Strenuous Physical Exercise”. In: *Journal of Proteome Research* 8.6 (June 2009), pp. 2966–2977. ISSN: 1535-3907. DOI: 10.1021/pr900081q. URL: http://dx.doi.org/10.1021/pr900081q.

[146] E. Chorell et al. “Physical fitness level is reflected by alterations in the human plasma metabolome”. In: *Molecular BioSystems* 8.4 (2012), p. 1187. DOI: 10.1039/c2mb05428k.

[147] R. Lehmann et al. “Medium chain acylcarnitines dominate the metabolite pattern in humans under moderate intensity exercise and support lipid oxidation”. In: *PLoS ONE* 5.7 (2010), e11519. DOI: 10.1371/journal.pone.0011519.

[148] A. K. Smilde et al. “Dynamic metabolomic data analysis: A tutorial review”. In: *Metabolomics* 6.1 (2010), pp. 3–17. DOI: 10.1007/s11306-009-0191-1.

[149] M. B. Eisen et al. “Cluster analysis and display of genome-wide expression patterns.” In: *Proceedings of the National Academy of Sciences of the United States of America* 95.25 (1998), pp. 14863–14868. DOI: 10.1073/pnas.95.25.14863.

[150] J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm". In: *Applied Statistics* 28.1 (1979), p. 100. DOI: 10.2307/2346830.

[151] L. J. Heyer, S. Kruglyak, and S. Yooseph. "Exploring expression data identification and analysis of coexpressed genes". In: *Genome Research* 9.11 (1999), pp. 1106–1115. DOI: 10.1101/gr.9.11.1106.

[152] M. Netzer et al. "Profiling the human response to physical exercise: a computational strategy for the identification and kinetic analysis of metabolic biomarkers". In: *Journal of Clinical Bioinformatics* 1.1 (2011), p. 34. DOI: 10.1186/2043-9113-1-34.

[153] E. Pohjanen et al. "A multivariate screening strategy for investigating metabolic effects of strenuous physical exercise in human serum". In: *Journal of Proteome Research* 6.6 (2007), pp. 2113–2120. DOI: 10.1021/pr070007g.

[154] A. Pechlivanis et al. "1H NMR-based metabonomic investigation of the effect of two different exercise sessions on the metabolic fingerprint of human urine". In: *Journal of Proteome Research* 9.12 (2010), pp. 6405–6416. DOI: 10.1021/pr100684t.

[155] A. Pechlivanis et al. "1H NMR study on the short-and long-term impact of two training programs of sprint running on the metabolic fingerprint of human serum". In: *Journal of Proteome Research* 12.1 (2013), pp. 470–480. DOI: 10.1021/pr300846x.

[156] B. Yan et al. "Metabolomic investigation into variation of endogenous metabolites in professional athletes subject to strength-endurance training." In: *Journal of applied physiology (Bethesda, Md. : 1985)* 106.2 (2009), pp. 531–538. DOI: 10.1152/japplphysiol.90816.2008.

[157] T. Jewison et al. "SMPDB 2.0: Big improvements to the small molecule pathway database". In: *Nucleic Acids Research* 42 (2014). DOI: 10.1093/nar/gkt1067.

[158] J. Hansen et al. "Exercise induces a marked increase in plasma follistatin: Evidence that follistatin is a contraction-induced hepatokine". In: *Endocrinology* 152.1 (2011), pp. 164–171. DOI: 10.1210/en.2010-0868.

[159] T. C. A. Akerstrom et al. "Oral glucose ingestion attenuates exercise-induced activation of 5'-AMP-activated protein kinase in human skeletal muscle". In: *Biochemical and Biophysical Research Communications* 342.3 (2006), pp. 949–955. DOI: 10.1016/j.bbrc.2006.02.057.

[160] A. K. Smilde et al. "Fusion of mass spectrometry-based metabolomics data". In: *Analytical Chemistry* 77.20 (2005), pp. 6729–6736. DOI: 10.1021/ac051080y.

[161] C. L. Kien. "Digestion, absorption, and fermentation of carbohydrates in the newborn." In: *Clinics in perinatology* 23.2 (1996), pp. 211–228. DOI: 10.1152/ajpendo.90748.2008.

[162] G. Biolo et al. "An abundant supply of amino acids enhances the metabolic effect of exercise on muscle protein." In: *The American journal of physiology* 273.1 (1997), E122–E129. URL: http://www.ncbi.nlm.nih.gov/pubmed/9252488.

[163] R. J. Havel, a. Naimark, and C. F. Borchgrevink. “Turnover rate and oxidation of free fatty acids of blood plasma in man during exercise: studies during continuous infusion of palmitate-1-C14.” In: *The Journal of clinical investigation* 42 (1963), pp. 1054–1063. DOI: 10.1172/JCI104791.

[164] V. Gaume et al. “Physical training decreases total plasma homocysteine and cysteine in middle-aged subjects”. In: *Annals of Nutrition and Metabolism* 49.2 (2005), pp. 125–131. DOI: 10.1159/000085536.

[165] J. R. Hoffman et al. “Effect of betaine supplementation on power performance and fatigue”. In: *Journal of the International Society of Sports Nutrition* 6.1 (2009), p. 7. DOI: 10.1186/1550-2783-6-7.

[166] Y. Yatabe et al. “Effects of Taurine Administration on Exercise”. English. In: *Taurine 7*. Ed. by J. Azuma, S. Schaffer, and T. Ito. Vol. 643. Advances in Experimental Medicine and Biology. Springer New York, 2009, pp. 245–252. ISBN: 978-0-387-75680-6. DOI: 10.1007/978-0-387-75681-3_25.

[167] J. E. Leklem and T. D. Shultz. “Increased plasma pyridoxal 5’-phosphate and vitamin B6 in male adolescents after a 4500-meter run”. In: *American Journal of Clinical Nutrition* 38.4 (1983), pp. 541–548. URL: http://www.ncbi.nlm.nih.gov/pubmed/6624696.

[168] H. Galbo. “Glucagon to graded and plasma and prolonged catecholamine exercise responses in man”. In: *J appl Physiol* 38.1 (1975), pp. 70–76. URL: http://www.ncbi.nlm.nih.gov/pubmed/1110246.

[169] E. Blomstrand and B. Saltin. “Effect of muscle glycogen on glucose, lactate and amino acid metabolism during exercise and recovery in human subjects”. In: *The Journal of physiology* 514.1 (1999), pp. 293–302. DOI: 10.1111/j.1469-7793.1999.293af.x.

[170] J. Tuomilehto et al. “Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance.” In: *The New England journal of medicine* 344.18 (2001), pp. 1343–1350. DOI: 10.1056/NEJM200105033441801.

[171] Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. *Database of Single Nucleotide Polymorphisms (dbSNP). dbSNP accession: rs7903146*. Version (dbSNP Build ID: 142). 2014. URL: http://www.ncbi.nlm.nih.gov/SNP/.

[172] A. Helgason et al. “Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution.” In: *Nature genetics* 39.2 (2007), pp. 218–225. DOI: 10.1038/ng1960.

[173] E. Renström. “Impact of transcription factor 7-like 2 (TCF7L2) on pancreatic islet function and morphology in mice and men”. In: *Diabetologia* 55.10 (2012), pp. 2559–2561. DOI: 10.1007/s00125-012-2659-1.

[174] A. P. Gjesing et al. “Carriers of the TCF7L2 rs7903146 TT genotype have elevated levels of plasma glucose, serum proinsulin and plasma gastric inhibitory polypeptide (GIP) during a meal test”. In: *Diabetologia* 54.1 (2011), pp. 103–110. DOI: 10.1007/s00125-010-1940-4.

[175] V. Lyssenko et al. “Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes”. In: *Journal of Clinical Investigation* 117.8 (Aug. 2007), pp. 2155–2163. DOI: 10.1172/JCI30706.

[176] M. I. McCarthy, P. Rorsman, and A. L. Gloyn. “TCF7L2 and diabetes: A tale of two tissues, and of two species”. In: *Cell Metabolism* 17.2 (2013), pp. 157–159. DOI: 10.1016/j.cmet.2013.01.011.

[177] L. Shu et al. “Transcription factor 7-like 2 regulates β-cell survival and function in human pancreatic islets”. In: *Diabetes* 57.3 (2008), pp. 645–653. DOI: 10.2337/db07-0847.

[178] S. F. Boj et al. “Diabetes risk gene and wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand”. In: *Cell* 151 (2012), pp. 1595–1607. DOI: 10.1016/j.cell.2012.10.053.

[179] S. E. Ross et al. “Inhibition of adipogenesis by Wnt signaling.” In: *Science (New York, N.Y.)* 289.5481 (Aug. 2000), pp. 950–953. DOI: 10.1126/science.289.5481.950.

[180] W. Shao et al. “The wnt signaling pathway effector TCF7L2 controls gut and brain proglucagon gene expression and glucose homeostasis”. In: *Diabetes* 62.3 (2013), pp. 789–800. DOI: 10.2337/db12-0365.

[181] S. F. a. Grant. “Understanding the elusive mechanism of action of TCF7L2 in metabolism”. In: *Diabetes* 61.11 (Nov. 2012), pp. 2657–2658. DOI: 10.2337/db12-0891.

[182] J. Adamski. “Genome-wide association studies with metabolomics”. In: *Genome Medicine* 4.4 (2012), p. 34. DOI: 10.1186/gm333.

[183] C. Gieger et al. “Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum”. In: *PLoS Genetics* 4.11 (Nov. 2008). Ed. by G. Gibson, e1000282. DOI: 10.1371/journal.pgen.1000282.

[184] T. Illig et al. “A genome-wide perspective of genetic variation in human metabolism”. In: *Nature Genetics* 42.2 (Feb. 2010), pp. 137–141. DOI: 10.1038/ng.507.

[185] C. Thamer et al. “Reduced skeletal muscle oxygen uptake and reduced β-cell function: Two early abnormalities in normal glucose-tolerant offspring of patients with type 2 diabetes”. In: *Diabetes Care* 26.7 (2003), pp. 2126–2132. DOI: 10.2337/diacare.26.7.2126.

[186] Q. Huang et al. “Metabolic characterization of hepatocellular carcinoma using nontargeted tissue metabolomics”. In: *Cancer Research* 73.16 (Aug. 2013), pp. 4992–5002. DOI: 10.1158/0008-5472.CAN-13-0308.

[187] H. G. Gika et al. “Within-day reproducibility of an HPLC-MS-based method for metabonomic analysis: Application to human urine”. In: *Journal of Proteome Research* 6.8 (Aug. 2007), pp. 3291–3303. DOI: 10.1021/pr070183p.

[188] M. Matsuda and R. A. DeFronzo. “Insulin sensitivity indices obtained from oral glucose tolerance testing: Comparison with the euglycemic insulin clamp”. In: *Diabetes Care* 22.9 (Sept. 1999), pp. 1462–1470. DOI: 10.2337/diacare.22.9.1462.

[189] J. C. Levy, D. R. Matthews, and M. P. Hermans. “Correct homeostasis model assessment (HOMA) evaluation uses the computer program”. In: *Diabetes Care* 21.12 (1998), pp. 2191–2192. DOI: 10.2337/diacare.21.12.2191.

[190] The Oxford Centre for Diabetes, Endocrinology and Metabolism (DTU). *HOMA Calculator*. 2004. URL: http://www.dtu.ox.ac.uk/Homacalculator/index.php.

[191] N. G. Forouhi et al. “Differences in the prospective association between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes: The EPIC-InterAct case-cohort study”. In: *The Lancet Diabetes and Endocrinology* 2.10 (2014), pp. 810–818. DOI: 10.1016/S2213-8587(14)70146-9.

[192] G. Boden and G. I. Shulman. “Free fatty acids in obesity and type 2 diabetes: defining their role in the development of insulin resistance and beta-cell dysfunction.” In: *European journal of clinical investigation* 32 Suppl 3 (2002), pp. 14–23. DOI: 10.1046/j.1365-2362.32.s3.3.x.

[193] G. Boden. “Interaction between free fatty acids and glucose metabolism.” In: *Current opinion in clinical nutrition and metabolic care* 5.5 (Sept. 2002), pp. 545–549. DOI: 10.1097/00075197-200209000-00014.

[194] G. Boden. “Obesity and diabetes mellitus–how are they linked?” In: *West Indian Medical Journal* 51 Suppl 1 (2002), pp. 51–54. URL: http://www.ncbi.nlm.nih.gov/pubmed/12050976.

[195] R. S. Rittmaster et al. “Androstanediol glucuronide isomers in normal men and women and in men infused with labeled dihydrotestosterone”. In: *Journal of Clinical Endocrinology and Metabolism* 66.1 (1988), pp. 212–216. DOI: 10.1210/jcem-66-1-212.

[196] Thermo Fisher Scientific. *SIEVE*. Version 1.2. URL: http://www.thermoscientific.com.

[197] D. Carr, N. Lewin-Koh, and M. Maechler. *Hexagonal Binning Routines*. 2010. URL: http://cran.r-project.org/web/packages/hexbin/.