An interactive online software platform for the analysis of small molecules using hyphenated mass spectrometry: MeltDB and ALLocator

Zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften an der Technischen Fakultät der Universität Bielefeld vorgelegte Dissertation

von

Nikolas Kessler

April 2018

Supervisors:
Prof. Dr. Tim W. Nattkemper
Prof. Dr. Alexander Goesmann

1st Referee:
Prof. Dr. Tim W. Nattkemper

Nikolas Kessler
Kulenkampffallee 79
28213 Bremen

Printed on non-aging paper according to DIN-ISO 9706.

Acknowledgments

I wish to take this chance to thank all the people who helped me during my studies and my PhD thesis, or even made them possible in one or another way.

At first, I wholeheartedly thank **my family**, for their trust and loving support throughout all the years.

I am very grateful to **Prof. Dr. Tim W. Nattkemper**, who many years ago asked me to join his workgroup, now called the **Biodata Mining Group**. Since then he supported me with his experience, guidance, and critical discussions in so many bioinformatics projects, including this PhD thesis.

I want to thank **Prof. Dr. Alexander Goesmann** for all his support and not least for giving me the chance to develop MeltDB 2.0 and ALLocator in company of the members of the **Computational Genomics** group and the **Bioinformatics Resource Facility**.

Very special thanks go to **Dr. Heiko Neuweger**, originally from this group, who not only allowed me to build upon his great MeltDB project, but who always helped me, supported me, and pushed my work - and he still does.

I also want to express my gratitude to **Prof. Dr. Karsten Niehaus**, **Prof. Dr. Thomas Noll**, **Prof. Dr. Jörn Kalinowksi**, and the members of their respective workgroups, who to a large extent shaped MeltDB and ALLocator with their experience, expertise, and constructive feedback.

Among them I want to additionally emphasize my friends and collaboration partners **Dr. Frederik Walter**, **Dr. Anja Bonte** and **Dr. Karin Gorzolka**. It was a pleasure and also fruitful to discuss with you both biological research and software development.

A big thank you goes to the **CLIB Graduate Cluster** and everyone involved; Not only for funding me and my PhD project, but also for the great environment built up by so many inspiring people. Many of them I call my friends and many of them were somehow involved in my projects as well.

Last but not least I thank all **my friends** who made this an incredibly great time, both inside and outside the university.

Nikolas Kessler, Bremen, April 2018

Contents

Acronyms xi

Glossary xiii

List of publications xv

1 Introduction 1

2 Background 5
2.1 Metabolome analysis . . . 5
2.2 Metabolites and their diverse characteristics . . . 7
2.3 Chromatography and mass spectrometry for metabolome analysis 9
2.3.1 Sample preparation for metabolome analysis . . . 10
2.3.2 Basics of chromatography . . . 11
2.3.3 Liquid chromatographs . . . 17
2.3.4 Gas chromatographs . . . 19
2.3.5 Basics of mass spectrometry . . . 21
2.3.6 Ionization methods . . . 22
2.3.7 Mass Analyzers . . . 26
2.4 Preprocessing of chromatography-hyphenated MS data . . . 28
2.4.1 Structure of chromatography-hyphenated MS data . . . 28
2.4.2 Mass spectral data preprocessing steps . . . 30
2.4.3 Chromatogram alignment . . . 31
2.4.4 Peak detection and quantitation . . . 31
2.5 Integration of metabolomics chromatographic data . . . 32
2.5.1 Spectra deconvolution . . . 33
2.5.2 Spectra matching . . . 34
2.5.3 Mass decomposition . . . 36
2.5.4 Metabolite identification . . . 37
2.5.5 Preparation of quantitation tables . . . 41
2.6 Analytical approaches to the metabolome . . . 42
2.6.1 Targeted analyses . . . 42
2.6.2 Metabolic profiling . . . 42
2.6.3 Untargeted analyses (metabolomics) . . . 43
2.6.4 Fingerprinting . . . 43

2.6.5 Stable isotope labeling 44

3 Current progress in metabolomics data analysis 47
3.1 Command-line tools 47
3.1.1 XCMS 47
3.1.2 CAMERA 48
3.1.3 MET-COFEA 48
3.1.4 X^{13}CMS 48
3.1.5 AMDORAP 48
3.1.6 apLCMS 49
3.1.7 mzMatch(-ISO) 49
3.2 Desktop applications 49
3.2.1 mzMine 49
3.2.2 MetAlign 50
3.2.3 Decon2LS 50
3.2.4 PeakML Viewer 50
3.2.5 MetExtract 50
3.3 Web applications 51
3.3.1 MetaboAnalyst 2.0 51
3.3.2 XCMSOnline 51
3.3.3 MetabolomeExpress 51
3.3.4 MetiTree 52
3.3.5 metaP-Server 52
3.3.6 MetFrag 52
3.3.7 ChromA 52
3.4 Public resources and repositories for metabolome informatics . . . 52
3.4.1 Kyoto Encyclopedia of Genes and Genomes 53
3.4.2 Metlin 53
3.4.3 Human Metabolome Database 53
3.4.4 National Institute of Standards and Technology 53
3.4.5 GOLM Metabolite Database 54
3.4.6 ChemSpider 54
3.4.7 MassBank 54
3.4.8 FiehnLib 54
3.4.9 MetaboLights 54
3.4.10 Metabolomics Workbench 55
3.4.11 PredRet 55
3.5 Summary 55

4 Challenges for computational metabolomics 57

5 ALLocator 1.0 **61**
5.1 System design and implementation . . . 63
5.1.1 System integration . . . 63
5.1.2 Data model . . . 66
5.1.3 User management . . . 66
5.2 Pre-processing methods . . . 69
5.3 Spectra deconvolution algorithm . . . 69
5.4 Metabolite annotation . . . 72
5.4.1 Manual annotation . . . 72
5.4.2 Search by monoisotopic mass (KEGG) . . . 74
5.4.3 Search by spectrum (MassBank) . . . 74
5.4.4 Mass decomposition and search by molecular formula (ChemSpider) . . . 76
5.4.5 Custom reference lists . . . 80
5.5 Data curation . . . 82
5.6 Summary . . . 86

6 STUDY: Amino-acid profiling in *C. glutamicum* strains **87**
6.1 Annotation of large neutral losses allows identification of ($\gamma-$)glutamyl dipeptides . . . 89
6.2 Data export and relative quantitation of arginine biosynthesis intermediates . . . 92
6.3 Discussion . . . 93

7 MeltDB 2.0 **99**
7.1 System design and implementation . . . 100
7.1.1 System integration . . . 101
7.1.2 System design and data model . . . 101
7.1.3 User management . . . 102
7.2 General workflow and integrated features . . . 102
7.2.1 Methods for data preprocessing . . . 102
7.2.2 Profiling methods for data integration . . . 104
7.2.3 User interfaces for all levels of data abstraction . . . 105
7.2.4 Statistics and data mining . . . 107
7.3 Summary . . . 110

8 STUDY: Multivariate GC-MS wheat data analysis **113**
8.1 Wheat sample preparation and GC-MS analysis . . . 115
8.2 Data preprocessing and feature annotation . . . 116
8.2.1 Unsupervised learning / dimensional reduction . . . 116
8.2.2 Supervised learning / classification . . . 117

8.3 Results 118
8.3.1 Unsupervised learning / dimensional reduction 118
8.3.2 Supervised learning / classification 125
8.4 Discussion 126

9 Discussion and Conclusion **129**

10 Contributions to computational metabolomics **133**

11 Bibliography **137**

Acronyms

m/z	Mass-to-charge ratio	9, 26–34, 42–44, 61, 70, 80, 82–84, 88, *Glossary:* m/z
Å	Ångström	8, *Glossary:* Å
API	application programming interface	53, 54
CAWG	chemical analysis work group	39
Da	Dalton	5, 8, 24, 25, 28, 36, 37, 44, 45, 70, 73, 76, 78, 79, 87–89, *Glossary:* Da
DRCC	Data Repository and Coordinating Center	55
EDA	exploratory data analysis	60
EI	electron ionization	9, 22, 23, 28, 33, 53, 54
EIC	extracted ion chromatogram	83, 85
ESI	electrospray ionization	9, 22, 23, 25, 33, 48, 61
GC	gas chromatography	9, 11, 12, 17, 19, 21–23, 28, 34, 62
GC-EI-MS	gas chromatography coupled to mass spectrometry via electron impact ion source	23, 24, 26, 59
GC-MS	gas chromatography coupled to mass spectrometry	1–3, 7, 21, 22, 33, 36, 42, 43, 47, 50, 51, 54, 55, 61, 101, 111, 131
GMD	Golm Metabolome Database	24, 54, 104
HMDB	Human Metabolome Database	5, 52, 53
IDEC	interactive data exploration and curation	55, 60, 63, 99, 102, 105, 129
InfoViz	information visualization	60

IUPAC	International Union of Pure and Applied Chemistry	21
KEGG	Kyoto Encyclopedia of Genes and Genomes	38, 52, 53, 65, 72–75, 77, 86, 87, 89, 97, 106
LC	liquid chromatography	9, 11, 12, 17, 18, 21–23, 25, 34, 59
LC-ESI-MS	liquid chromatography coupled to mass spectrometry via electrospray ionization	2, 3, 23, 25, 33, 34, 37, 40, 45, 59, 61–63, 69, 72, 74, 82, 129, 130, 133
LC-MS	liquid chromatography coupled to mass spectrometry	1, 2, 7, 33, 36, 42–44, 47–51, 55, 61, 62
MS	mass spectrometry	21–23, 26, 28, 30, 36, 37, 40, 41, 43, 44, 47, 52, 53, 61, 130
MS/MS	two-dimensional mass spectrometry	28, 40, 53, 65, 74, 75, 130
MS^n	multi stage mass spectrometry	52, 53
MSI	metabolomics standards initiative	39, 40
NIH	National Institute of Health	55
NIST	National Institute of Standards and Technology	53, 104
NMR	nuclear magnetic resonance	22, 37, 51–53
RI	Retention Index	53, 54
RT	Retention time	9, 12, 16, 31, 42, 43, 69–71, 82, *Glossary:* RT
SIL	stable isotope labeling	2, 3, 42, 45, 49, 50, 55, 57, 62, 80, 83, 86, 129, 134

Glossary

Å	Ångström, length unit, $1\ \mathring{A} = 100\ pm = 10^{-10}\ m$	8
Da	Dalton, atomic mass unit, $12\ Da = m_{Carbon}$	5
K	Distribution coefficient, the ratio of concentrations of a compound in a mixture of two immiscible phases at equilibrium.	12
m/z	The mass-to-charge ratio (sometimes mass-over-charge ratio) of a molecule is its exact mass divided by the number of charges this molecule carries.	26
pK_a	The pH at which an equal number of the functional groups of a compound are protonated or not.	8
RT	The time a compound remains in a chromatography column from injection to elution.	12

List of publications

Persicke M, Rückert C, Plassmeier J, Stutz LJ, Kessler N, Kalinowski J, Goesmann A, Neuweger H (2011) **MSEA: metabolite set enrichment analysis in the MeltDB metabolomics software platform: metabolic profiling of Corynebacterium glutamicum as an example.** Metabolomics 8: 310-322.

Gorzolka K, Lissel M, Kessler N, Loch-Ahring S, Niehaus K (2012) **Metabolite fingerprinting of barley whole seeds, endosperms, and embryos during industrial malting.** J Biotechnol 159: 177-187.

Kessler N, Neuweger H, Bonte A, Langenkämper G, Niehaus K, Nattkemper TW, Goesmann A (2013) **MeltDB 2.0-advances of the metabolomics software system.** Bioinformatics 29: 2452-2459.

Bonte A, Kessler N, Nattkemper TW, Goesmann A, Thonar C, Mäder P, Niehaus K, Langenkämper G (2013) **Einsatz von Protein- und Metabolit-Profiling-Methoden zur Unterscheidung von ökologischem und konventionellem Weizen.** Beiträge zur 12. Wissenschaftstagung Ökologischer Landbau, Bonn, 5. - 8. März 2013. Berlin: D. Neuhoff, C. Stumm, S. Ziegler, G. Rahmann, U. Hamm & U. Köpke. 362-365.

Kessler N, Walter F, Persicke M, Albaum SP, Kalinowski J, Goesmann A, Niehaus K, Nattkemper TW (2014) **ALLocator: An Interactive Web Platform for the Analysis of Metabolomic LC-ESI-MS Datasets, Enabling Semi-Automated, User-Revised Compound Annotation and Mass Isotopomer Ratio Analysis.** PLoS One 9: e113909.

Kessler N, Bonte A, Albaum SP, Mäder P, Messmer M, Goesmann A, Niehaus K, Langenkämper G, Nattkemper TW (2015) **Learning to Classify Organic and Conventional Wheat - A Machine Learning Driven Approach Using the MeltDB 2.0 Metabolomics Analysis Platform.** Front Bioeng Biotechnol 3: 1-10.

1 Introduction

In 1995 it was the first time that a full genome, the one of *Haemophilus influenzae* (Fleischmann *et al.*), was sequenced. The first eukaryotic genome followed in 1996 (*Saccharomyces cerevisiae*, Gofieau *et al.*). These breakthroughs in the field of genomics brought an unknown wealth of data. The inventory of genes and their DNA sequences could be used by genome researchers to link certain phenotypes and diseases to specific genes. But that wealth of data also posed new questions in the same order of magnitude, as a plethora of hypothetical genes and genes of unknown functions were discovered. To elucidate the roles, functions and interactions of all genes, all layers of the molecular network in a cell had to be explored. This initiated the post-genomic era and opened the doors for the research fields *functional genomics* and *systems biology*. New (high-throughput) technologies came up to explore the gene's products, the set of transcripts and proteins in a cell at a certain time, and bioinformatics had to develop in the same pace to provide the necessary tools for data analysis and evaluation.

As the youngest member of the main *Omes*, the metabolome was defined as the set of metabolites, i.e. the intermediate products of biochemical pathways, in a cell at a certain time (Oliver *et al.*, 1998). Unlike the constant genome, the transcriptome, proteome and metabolome are dynamic in nature. The quantities of their compounds are influenced by external and internal conditions, and they are regulated by each other. Metabolites introduce another unique feature, as they are not built up by a linear sequence of nucleotides (like genes) or amino acids (like proteins). They can only be identified unambiguously, if their complete three-dimensional structure of atoms is determined.

The metabolomics research field thus aims to investigate the complete set of metabolites in the individuals of a population in a systematic, but untargeted fashion. The term 'metabolic profiling' desrcibes a similar approach, which is tailored to a certain class of metabolites only. However, there is not a single technology which can cover the entire metabolome. Among the most popular instrumental setups for metabolome analyses are liquid chromatography coupled to mass spectrometry (LC-MS) and gas chromatography coupled to mass spectrometry (GC-MS).

The metabolomics research field involves a variety of scientific disciplines from physics and engineering, to chemistry and biology, to mathematics and statistics. Its integration into the *Omics cascade* additionally emphasizes the interdisciplinary

character. The same is reflected in the requirements for metabolomics informatics. The process from raw measurement data to the finally derived gain in biological knowledge often integrates cheminformatics, bioinformatics, statistics, data mining and visualization.

Similarly, metabolomics finds its application in various other disciplines. To mention only a few, these might be: Biotechnology, where it guides metabolic engineering; Food authenticity and quality control, where it helps to assess the origins, treatments or compositions of samples; or personalized medicine, where it helps to assemble panels of biomarkers for disease diagnosis and therapy selection.

Thanks to tremendous advances in analytical technologies, it became feasible to monitor many analytes at once, instead of only measuring one (German *et al.*, 2005). But unfortunately there is no single technology that is capable of covering the entire metablome. Thus ideally, the application, or the concrete scientific question posed to the metabolome, decides which instrumental setup will be used for the assessment of the metabolites. Two of the most popular combinations are LC-MS and GC-MS. Both cover different subsets of the metabolome, even though there is a large overlap. These technologies are considered comparably sensitive and they have their strengths in the identification of previously known compounds, but they are not quantitative methods *per se*. They can be leveraged to make quantitative statements though, by including internal standards of metabolites that are e.g. enriched in stable isotopes and thus produce separate but comparable signals. Such studies are referred to as stable isotope labeling (SIL) experiments.

The work presented in this thesis aims for the establishment of an interactive online software platform for the analysis of small molecules using GC-MS and LC-MS technology. To this end, two online platforms have been developed or extended: ALLocator was developed for the in-depth analysis of liquid chromatography coupled to mass spectrometry via electrospray ionization (LC-ESI-MS) experiment data, also introducing a novel algorithm for the processing of data from SIL experiments and thus enabling quantitation. Additionally, the MeltDB software (Neuweger *et al.*, 2008) was extended to facilitate the analysis of large sample cohorts and to add new features for the training and application of machine-learning classifiers.

In the upcoming background chapter 2 the reader is at first introduced to the topic of metabolome analysis and metabolites themselves. The following section 2.3 is dedicated to thoroughly explain the analytical methods chromatography and mass spectrometry, always keeping a special focus on aspects that are relevant for the preprocessing of the raw data (section 2.4) or for the integration and interpretation of the preprocessed data (section 2.5). Especially the latter section 2.5 explains the core principles that are foundation to the development of the ALLo-

cator software. The last background section 2.6 comprises the different analytical approaches to the metabolome and its assessment.

Chapter 3 presents a collection of existing software tools, resources, and repositories, which contribute in one or another way to metabolome informatics. The next chapter 4 shortly summarizes the main challenges for computational metabolomics along a workflow chart that lists the different tasks from raw data through to the final gain in knowledge about biology.

The ALLocator software platform, together with the new algorithm ALLocatorSD for the ion deconvolution of LC-ESI-MS spectra including SIL, are presented in chapter 5. The chapter comprises technical details like system design and implementation, but also a summary of the preprocessing methods and an in-depth view into the ALLocatorSD algorithm, as well as strategies for the interpretation, integration, and curation of the processed data and how the user is aided through the interactive graphical user interface.

The MeltDB 2.0 software platform and its advancements are covered in chapter 7. Similar to the previous chapter, the sections about MeltDB lead through technical details first, and then come to the features for preprocessing, data integration and data interpretation and how they are offered graphically.

Subsequent to these main chapters 5 'ALLocator 1.0' and 7 'MeltDB 2.0', there are summaries of metabolomics studies that were conducted using the respective software platforms. Chapter 6 describes an LC-ESI-MS SIL experiment analyzed within the ALLocator web platform to profile amino acids in different strains of *Corynebacterium glutamicum* and how the novel algorithm for ion deconvolution helped to identify ($\gamma-$)glutamyl dipeptides. The study in chapter 8 investigates the potential of metabolomics profiling techniques, bioinformatics and machine learning to distinguish organically grown wheat from conventionally grown wheat. For this, more than 300 GC-MS analyses from different batches were processed in MeltDB 2.0 and were used as training input for machine-learning classifiers.

A discussion and conclusion of the here presented progress in computational metabolomics is finally given in chapter 9. A reader who prefers to be guided through this thesis only along the sections which are most important for the main contributions of the work presented here may find a starting point in chapter 10.

2 Background

The first sections of this background chapter aim to comprise the foundations of the analytical approaches and the technologies applied for metabolomics research. In these sections the focus is always kept on those aspects that are most relevant for the data preprocessing, integration and interpretation techniques applied and presented in this thesis.

The next sections of this chapter address said data preprocessing and data integration topics. In the final section the different strategies to investigate small molecules or entire metabolomes are outlined.

2.1 Metabolome analysis

The *omics cascade* covers the investigation of the genome, the transcriptome, the proteome, and finally the metabolome (cmp. Table 2.1). The latter comprises the entire set of metabolites present in a cell or organism under a given set of conditions (Oliver *et al.*, 1998), while the definition of a metabolite is vague. The Human Metabolome Database (HMDB) for instance defines a metabolite as a small molecule with less than 1 500 Dalton (Da) that can be found at concentrations greater than $1\,\mu M$ in at least one condition and sample, and it should be of biological origin (Wishart *et al.*, 2007). But even this definition allows some exceptions, for example for less abundant metabolites of biomedical importance.

Accordingly, estimations for the number of metabolites that can be expected in a measurement, in an organism, or in an entire taxonomic kingdom vary greatly and are hard to compare. Nevertheless an overview shall be given: *Escherichia coli* are estimated to contain about 750 metabolites (Nobeli *et al.*, 2003), eukaryotes might well contain between 4 000 and 20 000 metabolites (Fernie *et al.*, 2004), and Fiehn (2002) expects up to 200 000 metabolites in the plant kingdom.

In any case, metabolomes are one layer integrated into the cascade of *omes* and need to be brought into context with the genome, transcriptome, and proteome. This is a complex task, as the interaction of these layers is not as straight-forward as often depicted in simplified schemes. The mRNA level for a given gene depends on the levels of transcription factors as well as of the activities of upsteram kinases and receptors. Next, the level of the respective protein is not only a function of the level of its mRNA, but additionally depends on the activity of the translational

Table 2.1: Key *omes*, their *omics*, and definitions.

Omes	**Omics**	**Ome Definitions**
Genome	Genomics	The complete nucleotide sequence in the genetic material of a living cell
Transcriptome	Transcriptomics	The complete set of mRNA present in the cell
Proteome	Proteomics	The complete set of proteins in the cell, including different post-translational modifications
Metabolome	Metabolomics	The complete set of all metabolites formed by the cell in association with its metabolism

apparatus, protein kinases, phosphatases, and proteases (Villas-Bôas *et al.*, 2007). In fact the correlations of expression levels of mRNAs and their corresponding proteins are disillusioningly low (Gygi *et al.*, 1999; Ideker *et al.*, 2001).

As metabolites are linked to each other in a network of anabolic and catabolic reactions (Weckwerth, 2003), any single molecule can be part of multiple pathways and have regulatory effects on a variety of biological processes. Less than 30 % of metabolites are part of only two reactions, while on the other hand more than 10 % contribute to more than ten reactions, and about 4 % are involved in twenty reactions or more (Förster *et al.*, 2003). Additionally, about two thirds of reactions in metabolic networks involve more than one substrate and one product (Nielsen and Oliver, 2005). Generally it is thus not possible to uniquely link a metabolite to a single genomic sequence (Bino *et al.*, 2004). But even though the inventory of metabolites can not be read from the genome directly, it is still the inventory of genes that defines which biochemical pathways are accessible. Consequently, the individual abundances of metabolites can be taken to characterize the biochemical response of a specific organism to a specific conditional perturbation.

Unlike metabolomics, not all metabolic analysis strategies aim to cover all (or as many as possible) small molecules, nor do all approaches aim for metabolite quantification and identification. According to Fiehn (2002) metabolomics has to be considered the 'quick-and-dirty' approach, trying to be as comprehensive and fast as possible, but not being ideally precise and reproducible in the determination of all metabolites. While advances in technology may eventually diminish the disadvantages, there will remain the trade-off between the accurate and quantitative analysis of a few metabolites versus the qualitative analysis of many metabolites (Nielsen and Oliver, 2005).

Table 2.2 lists and defines metabolomics and other analytical approaches to small molecules. A more detailed introduction to each is given in the last section of this chapter: 2.6 Analytical approaches to the metabolome.

The general approach towards the measurement of the metabolome of a complex sample is to first separate the small molecules by a chemical property, and then to resolve the gained fractions in some analyzer. Commonly applied technologies are mass spectrometers coupled to gas chromatography (GC-MS) or liquid chromatography (LC-MS). The next section outlines the chemical properties which are used by the diverse separation technologies to bring all the different small molecules in a complex sample apart.

Table 2.2: Definitions of analytical approaches towards the metabolome or parts of it.

Analytical approach	**Definition**
Metabolomics	Approaches to analyze the entire metabolome or a large fraction of it
Metabolic fingerprinting	Spectra from MS or NMR analysis provide a fingerprint that reflects the metabolites produced by a cell, without identification of specific metabolites
Metabolic footprinting	Like metabolic fingerprinting but targets the exometabolome - sometimes includes metabolite identification
Metabolic profiling	Approaches to analyze a certain group of metabolites, e.g. a class of small molecules like aminoacids - not necessarily quantitative
Metabolic target analysis	Quantitative analysis of selected metabolites, which are e.g. involved in a pathway of interest

2.2 Metabolites and their diverse characteristics

As described above, the metabolome is a complex compilation of very diverse small molecules. This diversity results from a variety of chemical and physical parameters that need to be taken into account before analysis, but also provide a chance to separate metabolites from each other, and finally to identify them: *Molecular weight*, *molecular size*, *polarity*, pK_a, *solubility*, *volatility*, and *stability* (Roessner, 2006).

The *molecular weight* of a metabolite can be read from its molecular formula, as

it is constituted by the weights of all atoms that form the molecule. By definition metabolites weigh below 1 500 Da (Wishart *et al.*, 2007). Rarely a molecular formula (and consequently a molecular weight) can be directly used to uniquely identify a metabolite. Different metabolites with the same molecular formula may feature different spatial arrangements of their atoms. These metabolites with different tridimensional structures are so-called isomers of each other. In addition to that, higher masses can be explained with multiple molecular formulas. The number of molecular formulas that have to be considered for a given mass rises with the mass and its inaccuracy.

The *molecular size* of a molecule is depending on its tridimensional structure and surrounding molecules that are noncovalently bound to it, like water. The resulting efficient volume is measured in Ångström (Å)^3.

The *polarity* of a molecule can be understood as its tendency to build noncovalent, polar bonds to other molecules. A ubiquitous example of a polar bond is the hydrogen bond, where a hydrogen attracts to a highly electronegative atom like oxygen or nitrogen of another molecule or intramolecular. More generalized it is the interaction of the positive end of one dipole to the negative end of another dipole, which requires atoms of different electronegativity. The polarity of a molecule closely correlates with its solubility in water (formation of hydrogen bonds with water) and its boiling point (formation of intermolecular polar bonds). Highly polar molecules are more soluble in water and have higher boiling points. As many functional groups in organic compounds form dipoles, it is possible to rank the respective chemical classes by polarity (Roessner, 2006):

$Amide > Acid > Alcohol > Ketone \sim Aldehyde > Amine > Ester > Ether > Alkane$

The polarity of functional groups can change with pH, though. Amines for example have higher polarity at low pH, while acids get more polar at high pH, because they form ions (R-NH_3^+ and R-CO_2^- respectively).

This behavior is described by the pK_a. That is the pH at which an equal number of functional groups, acidic or alkaline, are protonated or not. In other words, at a pH above or below the pK_a the molecules are ionized or natural.

The *solubility* of a molecule is related to molecular size, polarity, pK_a, temperature and the solvent. The solubility of a solute is defined as 'the analytical composition of a saturated solution, expressed in terms of the proportion of a designated solute in a designated solvent' (Freiser and Nancollas, 1987). Generally, polar and ionic metabolites will solve in polar solvents, while unpolar solvents can dissolve unpolar metabolites.

Also bound to polarity is the *volatility* of molecules, which is their tendency to vaporize or sublime. Its vapor pressure, the pressure at which its gaseous phase and condensed (liquid or solid) phase are in thermodynamic equilibrium, depends on its boiling point and thus also its polarity.

The thermodynamic and kinetic *stability* of a molecule is its resistance to be changed via degradation or other reactions. Thermodynamically unstable metabolites have a more negative Gibbs free energy (ΔG) and will be mostly converted to something else at equilibrium. Metabolites that react very fast, are kinetically unstable (Roessner, 2006).

2.3 Chromatography and mass spectrometry for metabolome analysis

The simple rationale behind the idea of coupling chromatography to mass spectrometry, is to separate molecules from complex samples to the extend that distinct molecules produce distinct signals that can be measured. Both these parts of the system, as well as the technical interface to connect them, have been implemented in various ways. Combined, they can be used to create *snapshots* of parts of the metabolome of a given sample.

The most important (or most applied) techniques of chromatography in the context of metabolome analysis are liquid chromatography (LC) and gas chromatography (GC). These separate molecules according to their chemical properties, which results in characteristic retention times (RT) for each molecule: Some molecules will pass the chromatography faster than others.

To technically couple LC and GC to mass analyzers, the molecules have to be ionized. This is usually realized with electrospray ionization (ESI) sources or electron ionization (EI) sources, respectively.

Typical mass analyzers are for example quadrupole analyzers, time-of-flight analyzers, and iontrap analyzers. They separate ionized molecules by their mass-to-charge ratios (m/z) and forward them to the mass detector which, finally, translates the analytes abundances into digital signals.

There is no single combination of chromatographic method, ion source, and mass analyzer that is capable of detecting and analyzing *all* small molecules of a metabolome. The instrumentation of choice depends on the scientific question and consequently on the analytical approach to the metabolome (cmp. table 2.2). Furthermore, the proper application of any of these instrumentations requires proper sampling and sample preparation at first (Villas-Bôas *et al.*, 2006):

a) The metabolism in the sample must be stopped. If metabolic reactions continue during analysis, and biochemical processes keep altering the current set of metabolites, it is not possible to take a *snapshot* of the metabolome.

b) Metabolites have to be extracted from the cells and their biological matrices to make them accessible to analysis.

c) Steps a) and b) are virtually impossible to perform without losses. Thus it may be necessary to concentrate samples before measurement in order to heave low abundant metabolites over the analytical detection limit.

While sampling and sample preparation seems to be out of the scope of this thesis, it is such a crucial and limiting step in metabolome analysis that it may not be unmentioned.

This section aims to explain the principles of chromatography and mass spectrometry as well as their most important technical implementations. Sample preparation is essential to understand the limitations of the technique and thus is delineated beforehand.

2.3.1 Sample preparation for metabolome analysis

Metabolome analysis aims to obtain *snapshots* of the metabolome, which is a function of the targeted organisms enzymatic toolbox and given environmental conditions. However, unlike taking a photograph, taking a *snapshot* of all the metabolite levels at a specific time point is a relatively time consuming process. During that time biochemical processes keep working, which is especially troublesome, because the procedure of 'taking the *snapshot*' changes the environmental conditions of the organism to which its metabolome will react quickly (within a few seconds). While many secondary metabolites, the end products of side pathways, are quite stable, the turnover rates of primary metabolites can be very high. As a consequence metabolite levels change and are no more representative for the time point the entire data acquisition began. This is why quenching is an indispensable step that has to be done within a few seconds after harvesting cells from a culture flask: A rapid change of pH or temperature is induced in the cells of a harvested sample in order to stop their metabolism (Villas-Bôas *et al.*, 2006).

There are various quenching protocols reported, often with different advantages or disadvantages depending on the targeted organism. For microbial cells a common approach is quenching with cold methanol while keeping the sample temperature below -20 °C. After that, cells are separated from culture broth by centrifugation, then metabolite extraction is performed. However, among others for the Gram-positive bacterium *Corynebacterium glutamicum*, a significant loss of intracellular aminoacids can be observed during cold methanol quenching (Wittmann *et al.*, 2004; Bolten *et al.*, 2007). Here, quenching with liquid nitrogen (-196 °C) is preferred, but demands centrifugation to be performed before the quenching, which comes at an additional cost of about fifteen seconds. Liquid nitrogen based protocols are also the most commonly applied techniques for quenching in animal and plant cells.

After stopping the biochemical processes in the cells, the next step is metabolite

extraction. By definition intracellular metabolites are located inside the cells, and they are present in a wide range of concentrations. In addition to that, they feature diverse molecular characteristics as discussed above. The purpose of metabolite extraction is to break the cells' envelopes and to make the small molecules accessible for measurement. Again requirements vary for different cell types. While animal cells are basically protected by a single fluid-mosaic lipid bilayer, Gram-negative bacteria form a cell wall structure that consists of an inner cell membrane, a peptidoglycan layer, and an outer membrane. Gram-positive bacteria lack the outer membrane in favor of a thicker peptidoglycan layer. The cell walls of yeasts, filamentous fungi, and plants feature even stronger cell wall compositions. Cell disruption becomes increasingly difficult for cell types as they were named here: from animal to plant cells (Villas-Bôas *et al.*, 2006).

Metabolite extraction is further complicated by the diverse characteristics of metabolites. In this regard, they can be seen in three classes: polar, unpolar, and volatile. For example, the proper selection or combination of solvents may allow to extract a specific group of metabolites, or even to extract polar and unpolar molecules in one step. However, there is no single method that is capable of extracting all intracellular metabolites at once.

In total, it is virtually impossible to extract the entire set of present metabolites without chemical alterations and/or losses in concentration. The latter can be grave to the extend that certain metabolite concentrations are pushed below the detection limit. This can be mitigated with an additional step of sample concentration. In case of aqueous solutions for instance, water can be removed by lyophilization. However, lyophilization can also cause additional loss of certain metabolites, leading to a discrimination of these molecules.

After all, any available method for metabolite extraction is somehow biased (Weckwerth, 2003).

2.3.2 Basics of chromatography

The term 'chromatography' (from Greek *chroma* 'color' and *graphein* 'to write') was pinned by M. S. Tswett, who is regarded as the inventor of the technique, starting the development of the method in 1903 (Ettre and Sakodynskii, 1993a,b).

This section aims to explain the general principles of chromatography - or what chromatograms are and how they help to determine the molecular composition of a sample. These basics refer to both LC and GC.

Chromatography describes the use of a mobile phase (a liquid or a gas) that moves along a stationary phase in a column and transports the molecules of a sample. The stationary phase is usually a film that is chemically bound to the column surface and behaves like a liquid. Different molecules (M_1 and M_2) will 'favor' either the mobile phase or the stationary phase to a specific extend, which depends

on their molecular properties. This is represented by their distribution coefficients K_1 and K_2. Whenever the mobile phase moves to the next section of the column, molecules will migrate to the stationary phase according to their distribution coefficients. If the flow rate of the mobile phase is not too high, eventually equilibrium will be reached.

Molecules which are located in the stationary phase are retained in reference to the mobile phase, and thus they will elute (come off the column) later; molecules with a lower K have a higher retention time (RT). When the concentration of compounds in the eluting mobile phase at the end of the column is measured and plotted over time, a chromatogram is obtained. If, in the perfect case, the chromatography was able to completely separate the molecules M_1 and M_2, the chromatogram will display two separate peaks with no overlap. A perfect separation means that the signal at the end of the column drops down to the baseline between all peaks.

There are several factors that deteriorate the separation efficiency of a system (LC or GC) and broaden peak shapes, though. The most important of these are marked up in the following (Smedsgaard *et al.*, 2006):

Eddy diffusion Not all moieties of a compound take the same route through the column. Thus they have different retention times. The eddy diffusion only depends on the packing geometry of the column. A very uniform packing results in a very small eddy diffusion.

Longitudinal (or axial) diffusion The diffusion of the compounds along the axis of the column depends reciprocally on the flow rate. The faster the mobile phase flows through the column, the lower the longitudinal diffusion.

Resistance to mass transfer To a certain amount, and depending on both mobile and stationary phase, molecules can be hindered to move from one phase to the other. This effect prolongs the time it takes to reach the equilibrium of the compound concentrations in the mobile and the stationary phase at each section of the column. The impact on the peak shape increases with the flow rate.

The sum of these three dispersion effects, plotted against the mobile phase linear velocity u, is called a *van Deemter plot*. It shows the height equivalent of a theoretical plate H, which is used to measure the separation power of the system and can also be seen as the quotient of the column length L and the number of plates N (see equations (2.1) and (2.2)). The number of plates describes the number of equilibria a substance can reach between the mobile and stationary phase in a defined section of the column. A higher N (a lower H) results in more equilibria and consequently in a better separation (Smedsgaard *et al.*, 2006).

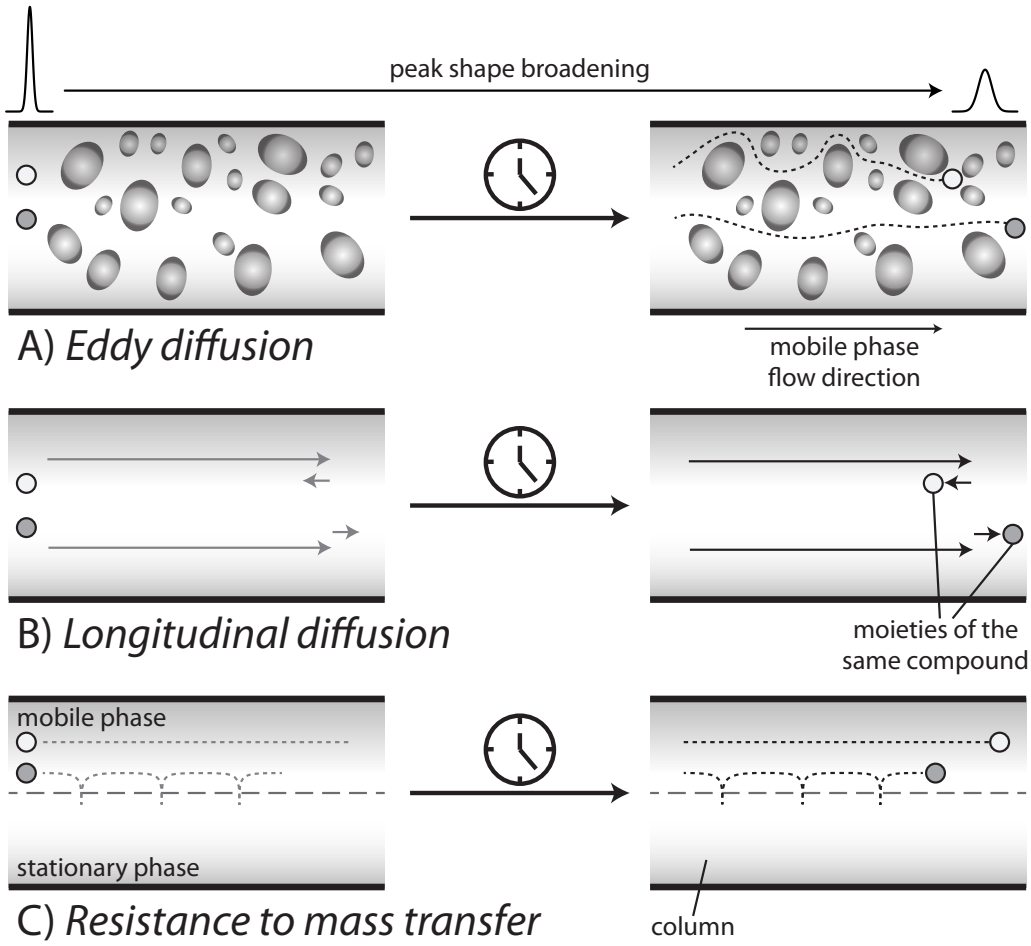

Figure 2.1: The three major dispersion effects and how they broaden the peak shape of eluting compounds. **A)** Eddy diffusion: Not all molecules take the same route through the packing material of the column; **B)** Longitudinal diffusion: Regardless of the flow direction, a small diffusion occurs into all directions, including the opposite direction. **C)** Resistance to mass transfer: The actual chromatographic effect - molecules need some time to migrate from one phase to the other.

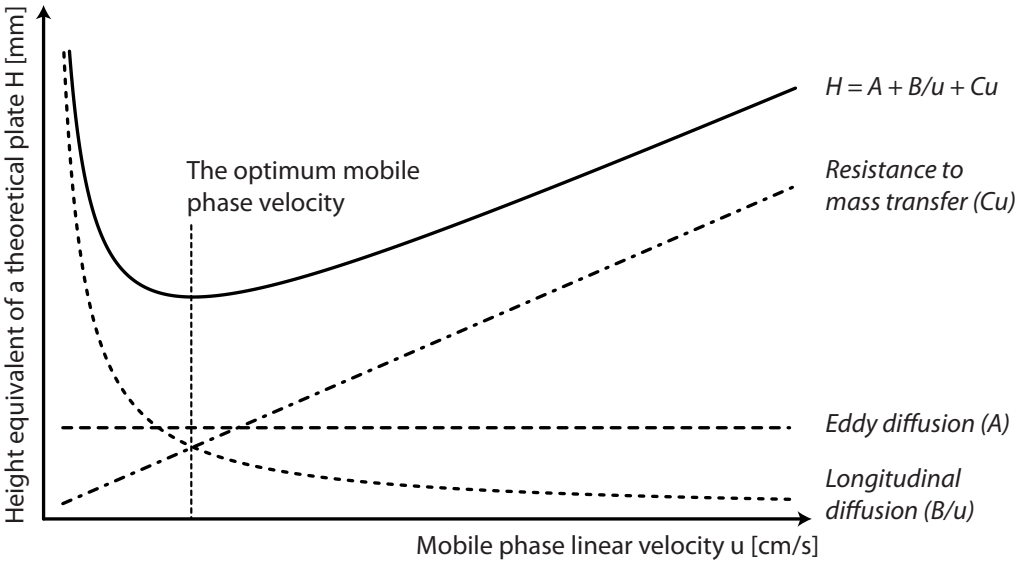

Figure 2.2: The van Deemter plot combines the three major dispersion effects to explain the optimum mobile phase linear velocity u.

$$H = A + B/u + Cu \tag{2.1}$$

H	equivalent of a theoretical plate height
A	eddy diffusion
B	longitudinal diffusion
C	resistance to mass transfer
u	mobile phase linear velocity

$$H = L/N \tag{2.2}$$

L	column length
N	number of plates

The van Deemter plot shows that there is an optimal u. Furthermore it appears that increasing u has a less severe effect on separation efficiency than decreasing u.

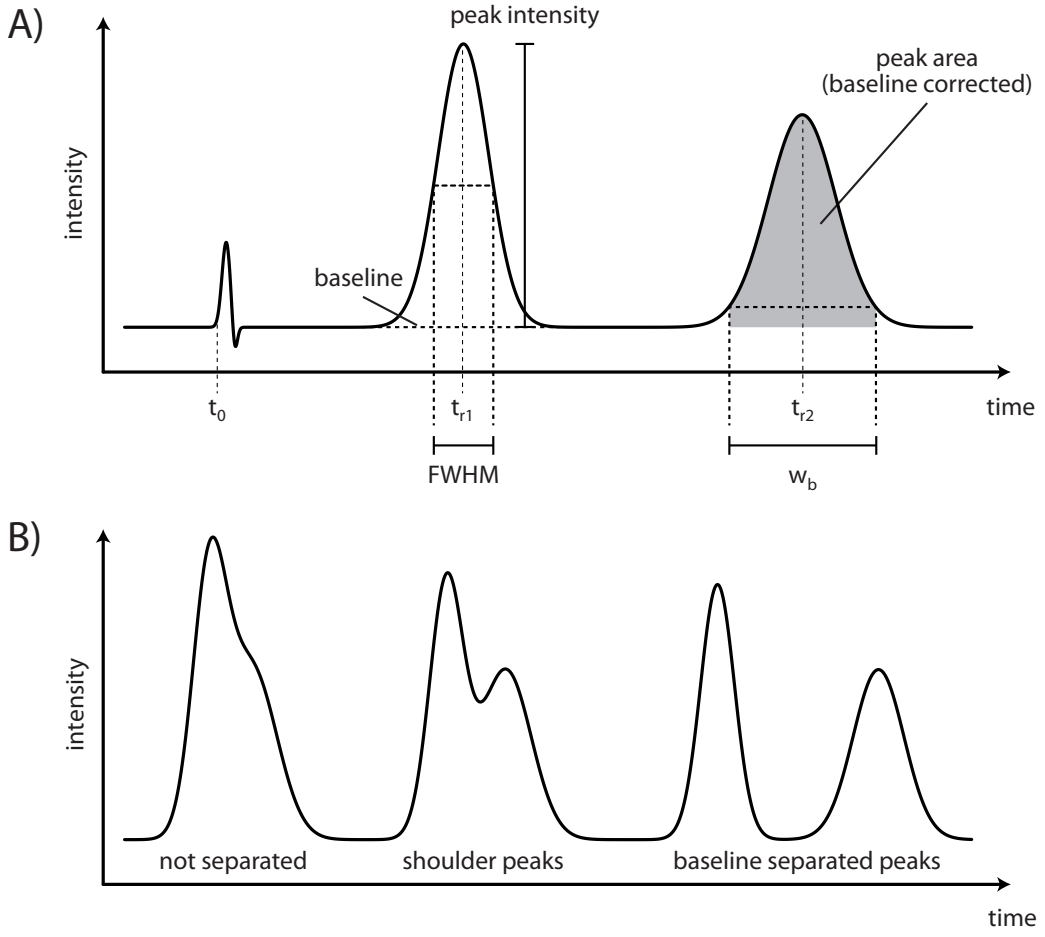

Figure 2.3: Parameters and overlapping shapes of chromatographic peaks. **A)** Schematic representation of peaks in a chromatogram and their parameters: t_0 is the dead time, the time that passes from sample injection to the first detected signal at the end of the colum; t_{r1} and t_{r2} are retention times of two eluting compounds; $FWHM$ is the *Full Width at Half Maximum* of a peak; w_b is the peak width at baseline. **B)** Schematic representation of peak forms that coeluting compounds may produce in a chromatogram.

The finally obtained chromatogram consists of the detector signal over time that is acquired at the end of the column. Recording starts with sample injection, and when a compound elutes, a signal peak is produced at that time - the compounds retention time RT. The quantity of that compound is inferred from either the peaks' height or the area under this peak. It is relatively simple to obtain the peak height: it is the signal intensity at the local maximum in the chromatogram minus the chromatogram base line (background noise). Obtaining the area under the peak however, requires to find the start and the end of a peak, which is difficult in noisy chromatograms and gets further complicated (or impossible) when peaks overlap. This again underlines the importance of a high separation efficiency.

The fraction of its retention time that a compound spent in the stationary phase is called its capacity factor k_r and is calculated as in equation (2.3). The selectivity α (see equation (2.4)) is the quotient of the capacity factors of two compounds. α can be applied to either compare the behavior of two different compounds in the same column, or to compare the behavior of one compound in two different columns (Smedsgaard *et al.*, 2006).

$$k_r = \frac{t_r - t_0}{t_r} \tag{2.3}$$

k_r capacity factor
t_r retention time
t_0 dead time - the time from sample injection to the first recorded signal

$$\alpha = \frac{k_{r2}}{k_{r1}} \tag{2.4}$$

α selectivity

To assess the resolution R with which the chromatographic system can separate two compounds C_1 and C_2, equation (2.5) is applied. Is the resolution larger than 1.5, the peaks of C_1 and C_2 are baseline separated and their peak areas can be analyzed independently.

$$R = 2 \cdot \frac{t_{r2} - t_{r1}}{w_{b2} + w_{b1}} \approx 1.18 \cdot \frac{t_{r2} - t_{r1}}{FWHM_2 + FWHM_1} \tag{2.5}$$

R	chromatographic resolution for two compounds C_1 and C_2
t_{r1}	retention time of C_1
w_{b1}	peak width at baseline of C_1
$FWHM_1$	*Full Width at Half Maximum* of the peak of C_1
	1.18 is approx. the ratio of $FWHM$ to w_b of a Gaussian curve.
	It can also be written as: $(2 \cdot \ln 2)^{1/2}$.

The principles presented in this section generally apply to both LC and GC. Nevertheless they differ very much in their technical characteristics and in the set of metabolites they can be applied for. These characteristics and their implications are addressed in the following sections.

2.3.3 Liquid chromatographs

The key concept of liquid chromatography is to transport a sample in a liquid mobile phase along the stationary phase in a separation column. A scheme of a liquid chromatograph and its parts is depicted in Figure 2.4.

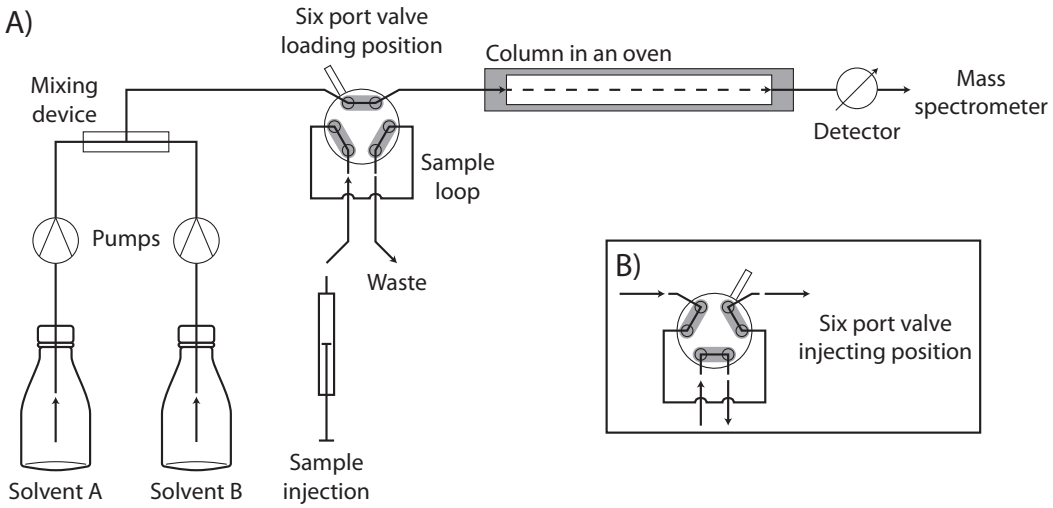

Figure 2.4: A) Exemplary scheme of a liquid chromatograph. Solvents A and B are pumped through the six port valve (sampler) into the separation column which is kept at temperature in an oven. Eluting compounds are detected (e.g. UV, fluorescence, or conductivity) at the end of the column and streamed into a mass spectrometer. The six port valve is depicted in loading position: The sample fills the sample loop; B) The six port valve in injecting position: The mobile phase stream injects the sample from the loop into the column.

Solvent reservoirs are connected to a six port valve via pumps. The valve then connects the solvent stream with the sample injection and the separation column, which can be kept at temperature in an oven. At the end of the column eluting compounds are detected and streamed into a mass spectrometer (if attached).

Two or more solvents constitute the mobile phase, and they are mixed at specific ratios that can be changed over time in order to create gradients. Applied solvents feature different 'eluting powers': changing the composition of the mobile phase changes its selectivity (see equation (2.4)). The amount each solvent contributes to the mobile phase composition is typically given in percentages. For each solvent percentages below 5 % are avoided though, because these are hard to dispatch accurately for technical reasons. Furthermore the pumps for the solvents are required to deliver a pulse-free flow at a precise flow rate and typically include a degasser to remove dissolved air (bubbles). The exact set up of these components varies and is shown only exemplary in Figure 2.4.

The mobile phase stream is led through a six port valve into the separation column. Of the six ports each is connected to exactly one neighboring port. This configuration can be rotated to connect each port to its other neighbor, instead. In the first configuration (loading position, Figure 2.4A) the mobile phase is directly streamed into the separation column. The sample however is injected into a loop of a defined length and diameter (and thus volume), which does not yet have contact to the mobile phase. Switching the configuration to the injecting position will lead the mobile phase through the sample loop, pushing the defined volume of the sample into the column (Figure 2.4B). The only critical parameter of this very reliable part of the system is the time it needs to transfer the sample to the column. That should never be above one second to limit peak broadening.

LC separation columns can be chosen from a very wide palette. Often the steel tubes are packed with porous particles made of silica or polymers. In case of silica particles, these can be applied for normal phase (NP) LC, making use of their polar silanol groups (Si-OH) in combination with an apolar mobile phase (e.g. Chloroform). In NP, polar compounds are retained most, while apolar compounds elute quickly. For reversed phase (RP) LC in contrast, the silanol groups are covered with apolar carbon chains like octyldecyl (C-18) chains. In the latter case a polar mobile phase is applied in a gradient from a very polar solvent (water) to a less polar solvent (e.g. acetontrile). RP retains mainly apolar compounds and is the most commonly used LC technique in metabolomics. Conceptually between NP and RP chromatography is the so called aqueous normal phase (ANP) chromatography. Here, the stationary phase features a hydride surface with Si-H groups instead of Si-OH. The mobile phase consists of a comparatively unpolar component (methanol or acetontrile) and a small amount of water. This results in some retention for both polar and apolar compounds. This is by far not a complete list of existing LC phase systems, but a more extensive enumeration exceeds the scope

of this thesis. It can already be seen though that the selection of stationary and mobile phases, as well as the design of mobile phase composition gradients give a lot of room for fine tuning and optimization.

Attached to the end of the column, a detector measures the amount of eluting compounds. As it is important to do nondestructive measurement if the columns effluent should be passed on to a mass spectrometer, UV and fluorescence detectors are often applied at this point. These however require molecules to have spectrometric features to detect them, which is not the case for many metabolites. This is not very critical though, as the attached MS is a very powerful detector itself as will be explained in section 2.3.5.

2.3.4 Gas chromatographs

Gas chromatographs are comparatively simpler than liquid chromatographs. Here the mobile phase usually consists of helium gas which is led through a 10 - 100 m long tubular to which the stationary phase is chemically bound. The main technical parts of a GC are a gas supply, an injector, the column in an oven, and finally a detector (see Figure 2.5A-C). Of these, the injector is the most difficult component with critical influence on peak shapes.

On the one hand the injector has to transfer the sample very rapidly into the carrier gas stream and onto the column, to avoid peak broadening. On the other hand it may be of utmost importance to transfer as much of the sample to the column as possible, in order to overcome detection limits. In modern GC practice this is realized with two different modes: *split* and *splitless* injection (see Figure 2.5B and C). These modes are a way to control the so-called split ratio, which describes the share of incoming gas flow that is actually transferred to the column. In split mode the split vent is open during sample injection. In this case the majority of both gas and the vaporized sample go to waste, and only about one to five percent enter the column. In splitless mode the split vent is closed at the beginning of sample injection, which results in almost zero sample loss. In the latter mode the transfer of the sample to the column can take several minutes and demands focusing the sample at column entrance, or peaks would be broadened to an unacceptable amount. To this end thermal and solvent focusing effects are applied, which counteract peak broadening. The column is situated in an oven to control its temperature, which can reach up to $250\,^{\circ}\mathrm{C}$. This temperature should be tailored to the analytes of interest, or to their boiling points and thermal stability.

Indeed, as described in section 2.2, the molecules boiling points and thermal stability are the characteristics which define, whether they can be analyzed via GC. To get the molecules into the column they need to be vaporized, which requires a certain volatility. Most metabolites with polar functional groups would thus be excluded from this analytical technique. To make them available for GC

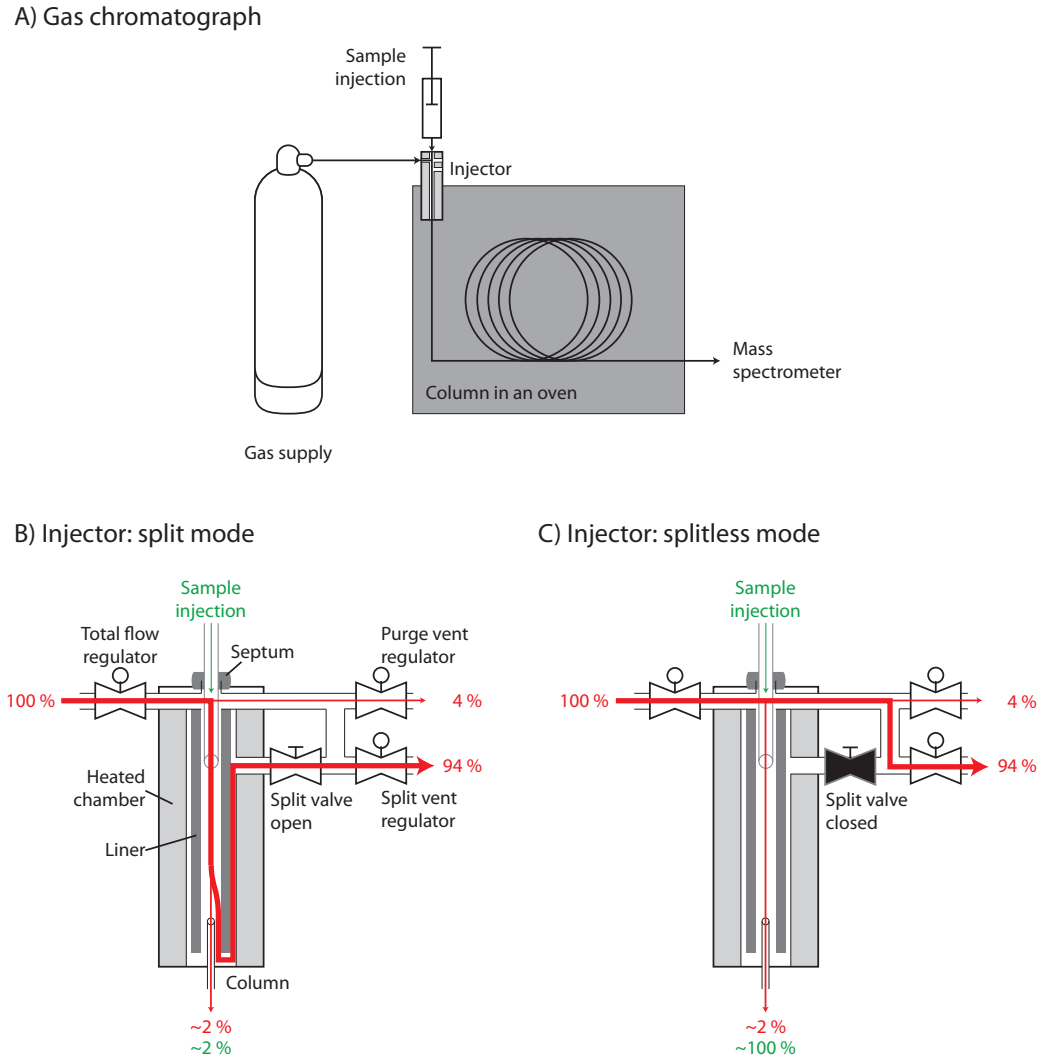

Figure 2.5: A) Exemplary scheme of a gas chromatograph. The carrier gas (helium or nitrogen) collects the sample in the injector and transfers it through the heated column. Finally, the separated sample gets into the mass spectrometer; B) Injector in split mode: The sample is transfered to the column rapidly but with a great loss; C) Injector in splitless mode: Almost the complete sample is transferred to the column at the cost of time and consequently peak broadening; In B) and C): Note that the injection needle is pushed into the liner for a certain length; The purge vent is to avoid septum bleeding and contaminations on the column; The shown flow rate percentages are examples and subject to fine tuning. Figure adapted and extended from Smedsgaard *et al.* (2006).

nevertheless, they have to be *derivatized*. Their functional groups are masked (substituted) with apolar trimethylsilyl (-Si(CH_3)$_3$) groups using e.g. *N*-methyl-*N*-trimethylsilyl-trifluoroacetamide (MSTFA). The resulting derivatives are apolar enough to be vaporized in the injectors heated chamber and to enter the column with the gas flow.

In the context of this thesis, GC is considered as a subpart of GC-MS. A mass spectrometer is attached to the end of the GC column to analyze the separated output. Keep in mind that most elutes that arrive in the MS are the volatile derivatives, not the original metabolites themselves.

2.3.5 Basics of mass spectrometry

Broken down to the essential, mass spectrometers apply electric and/or magnetic fields to determine mass-to-charge ratios (m/z $[Da/e]$) of ions in a gas phase and record their abundances. These capabilities make mass spectrometers very powerful analyzers which can be applied stand-alone or attached to an LC or GC for instance. In any way analytes need to be ionized and brought into the gas phase to make them detectable by mass spectrometry (MS). This is realized in the ion source - the first part of the MS which can be considered as some kind of adapter to the preceding chromatography instrument. As it is a fundamentally different request to ionize and bring the already gaseous eluents from a GC into the MS, than to ionize and evaporate the eluents from the mobile phase of an LC, these tasks are realized by different devices. The most common ion source for GC-MS is the rather simple electron impact (EI) ion source. Decades later a reliable solution was found to also couple LC to an MS: the electrospray ionization (ESI) ion source (Smedsgaard *et al.*, 2006).

The ionized analytes are led from the ion source into a device which determines their mass-to-charge (m/z) ratios. It is hence called the *mass analyzer*. There are various technical implementations of mass analyzers, but all require high vacuum and make use of electric and/or magnetic fields to determine the ions m/z values. The most important characteristics of mass analyzers are their mass accuracy and mass resolution. While mass accuracy is simply a measure for how exact mass measurements are (see equation (2.6)), there are many definitions for mass resolution in literature. The International Union of Pure and Applied Chemistry (IUPAC) defines *resolution in mass spectrometry* as in equation (2.7) (Murray *et al.*, 2013).

$$massaccuracy = \frac{m/z_{exp} - m/z_{theor}}{m/z_{theor} \cdot 10^6} \tag{2.6}$$

m/z_{exp} Experimentally measured mass-to-charge ratio
m/z_{theor} Theoretical mass-to-charge ratio

$$R = \frac{m/z}{\Delta m/z_{FWHM}} \tag{2.7}$$

$\Delta m/z_{FWHM}$ Peak width at 50 % of peak maximum

$$resolvingpower = \frac{m/z_1}{m/z_2 - m/z_1} \tag{2.8}$$

$m/z_1, m/z_2$ Two mass peaks of same height that overlap at a certain percentage (commonly 10 %) of peak height

Mass resolutions of analyzers are in the range from unit mass resolution (1:1000 to 1:2000) to as high as 1:100,000. Their mass accuracies can be as high as 1 ppm (parts per million) and even better, which eventually allows to unambiguously separate all feasible molecular formulas for masses below 1000 Da (cmp. section 2.5.3 on mass decomposition).

The actual m/z values of the ions are measured by detectors at last. Detectors measure the ion current or ion count as a function of time, which can be translated to m/z ratios according to the mass analyzers working principles. The detection relies on signal amplifiers and analog to digital converters which have to perform in high-speed to obtain the best possible mass resolution. Detectors are also key to sensitivity, which in MS is orders of magnitude higher than in nuclear magnetic resonance (NMR) (Scheubert *et al.*, 2013). The digitized signals are finally passed on to the mass spectrometers computer-assisted data system.

In the following, common implementations of the relevant parts are described in more detail: see sections Ionization methods (2.3.6) and Mass Analyzers (2.3.7).

2.3.6 Ionization methods

The purpose of the ion source is to transfer the sample from the mobile phase of the chromatograph into the gas phase, ionize it, and bring it into vacuum. In case of a classical GC-MS, which is connected via EI, the sample is already in the gas phase when it comes off the GC column. To couple an LC to an MS, the most commonly used ionization method is ESI. Both ionization methods are shortly outlined in this subsection, putting the focus on the respective fragmentation patterns and consequently on the mass spectra that are produced.

Electron impact (EI) ion source for GC-EI-MS

In an EI source, the compounds eluting from the GC column are immediately led into a crossing beam of electrons that are accelerated to typically 70 eV. The volume of the ion source is kept in a high vacuum ($< 5 \times 10^{-5}\ hPa$), such that the eluting compounds expand from the column violently, quickly increasing the distance between all molecules and thus avoiding molecule-molecule collisions and reactions. If one of the energetic electrons of the beam hits an eluting compound, another electron of the compound is 'kicked out', which first creates a positive-charged radical ion, and second brings excess energy into the molecule which in most cases leads to additional bond breakages: Fragmentation of the molecule occurs. The positive-charged ions are accelerated and transfered into the mass analyzer by another electric potential (Smedsgaard *et al.*, 2006).

Fragmentation patterns in EI are highly reproducible and very characteristic for their analytes. The spectra that are finally measured by the mass analyzer will mainly feature signals for the different fragments of the ionized compound and their isotopologues. Often the compound itself has been entirely fragmented due to the high excess energy, such that there is little or no signal left for its specific mass. This is why EI is often referred to as a 'hard ionization technique'. Additionally keep in mind that the compounds eluting from the GC column are typically derivatives of the original metabolites (see 2.3.4 Gas chromatographs). The high reproducibility of EI spectra allows to establish large spectral databases like NIST (NIST/EPA/NIH, 2014) or the GMD (Steinhauser *et al.*, 2004), which can be queried by virtually any gas chromatography coupled to mass spectrometry via electron impact ion source (GC-EI-MS) user to identify 'unknown knowns', i.e. to identify metabolites that have been identified somewhere else before (see 2.5.2 Spectra matching).

Electrospray ionization (ESI) ion source for LC-ESI-MS

The major challenge in coupling LC and MS is the fact that the compounds elute from a liquid mobile phase and must be brought into vacuum, before they can be transferred into the mass analyzer. Effectively, this became feasible with the advent of ESI, which made LC-ESI-MS highly popular and is the most widely applied technology in metabolomics mass spectrometry nowadays. The ESI process, which occurs at atmospheric pressure, is not yet fully understood. The mobile phase from the LC that carries the compounds in a solvent is pumped through a narrow steel needle, to which a certain voltage is applied. If that voltage reaches a specific threshold, the eluting liquid gets highly charged and forms a so-called Taylor cone at the end of the needle, getting thinner and thinner. Below a certain diameter at the tip of this cone the Rayleigh limit will be reached: The surface

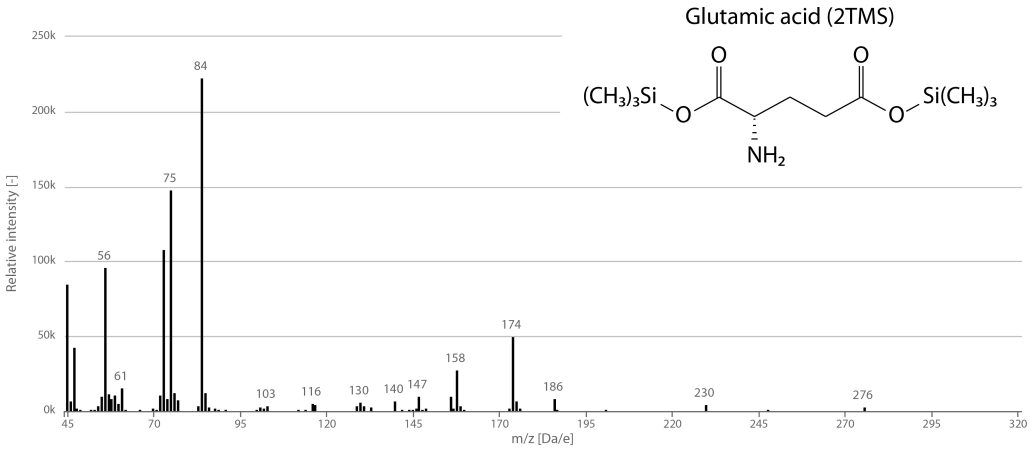

Figure 2.6: GC-EI-MS spectrum of the derivatized gluatmic acid (2TMS). The monoisotopic mass of this analyte is 291.13221 Da and its molecular formula is $C_{11}H_{25}NO_4Si_2$. The spectrum was taken from the GMD. While the figure was modified afterwards, no spectral information has been changed. Original authors of the spectrum: Boelling C, Liebig F, Erban A, Kopka J, Max Planck Institute of Molecular Plant Physiology, Department of Molecular Plant Physiology (Prof. Willmitzer L), Am Muehlenberg 1, D-14476 Golm, Germany.

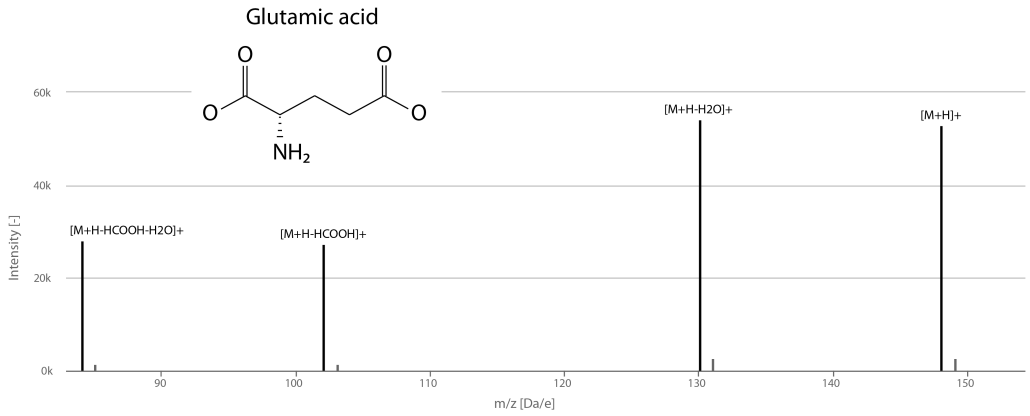

Figure 2.7: LC-ESI-MS spectrum of glutamic acid after deconvolution. Its monoisotopic mass is 147.05316 Da and its molecular formula is $C_5H_9NO_4$. The spectrum was taken from an ALLocator experiment to analyze complex samples of *C. glutamicum* measurements (Kessler *et al.*, 2014). While the figure was modified afterwards, no spectral information has been changed.

tension forces of the cone can no longer keep the high number of charges and it explodes into a number of droplets. Heated nitrogen gas is then used to make the solvent evaporate, such that these droplets get smaller and smaller again, eventually resulting in a Rayleigh limit-driven explosion into even smaller droplets. This process stops, when the droplets are small enough and contain few enough charges to hold them. Other theories are discussed, but exceed the scope of this work. In the end however, the very fine droplets are pumped into a high vacuum and get transferred into the mass analyzer (Smedsgaard *et al.*, 2006).

Of high interest in the context of this thesis is that ESI is a rather 'soft ionization technique'. The described process will mostly create $[M+H]^+$ or $[M+Na]^+$ ions in positive mode, or $[M-H]^-$ or $[M+Cl]^-$ ions in negative mode and rather few fragments (cmp. $[M+H-H_2O]^+$, $[M+H-CO_2]^+$, ...). Not less important is that ESI is very prone to its own parameterization, the solvent used in the LC method, and the complexity of the sample. So-called matrix effects may occur, where certain analytes that are easier ionized (i.e. they can carry charges easier) prevent other analytes from being ionized. This may result in a loss of signals of the latter - or even in a 'short-circuit' that mutes the signal altogether (Choi *et al.*, 2001; Sterner *et al.*, 2000; Annesley, 2003). Even though spectra are still highly reproducible and characteristic if all parameters are kept constant, the diversity of applied LC methods and ESI settings make the establishment of spectral databases for LC-ESI-MS

less applicable compared to GC-EI-MS spectral databases (Bino *et al.*, 2004). Fig. 2.7 presents a cleaned up (deconvoluted, see section 2.5.1) spectrum of glutamic acid as measured in a complex *C. glutamicum* sample (Kessler *et al.*, 2014).

2.3.7 Mass Analyzers

Mass analyzers, together with the attached detectors, perform the core tasks of mass spectrometry (MS): separating and assessing ions according to their mass-to-charge ratio (m/z) values. There is a variety of technologies available for this step, all of them make use of electric and/or magnetic fields and thus rely on the molecules' ionization beforehand. The performance of mass analyzers may be evaluated using the following characteristics of main interest: Mass range limit, mass accuracy, mass resolution, analysis speed, and sensitivity (de Hoffmann and Stroobant, 2001).

To finally record the abundances of analyzed ions, their amounts have to be converted from analog to digital. In the end, the digital output can be seen as pairs of m/z values and intensities, so-called profile spectra, which can eventually be mapped to retention times, if a chromatography preceded. These detection- and conversion-devices are referred to as 'detectors' in the following, but they are not discussed in detail in this thesis.

Time-of-flight mass analyzer

The time-of-flight (ToF) mass analyzer is very simple in principle: Ions are accelerated by an electric field and pushed into a vaccum tube - at the end of that tube a detector and a timer yield the abundance and the required time of ions to fly through the tube. The flight time t of an ion until it reaches the detector is proportional to the square-root of its m/z ratio. The scheme of a ToF is depicted in Fig. 2.8 a).

In a ToF, all ions arriving at the analyzer are measured in a single push event. Per second multiple ten thousands of push events occur and the resulting spectra are typically summed up to one spectrum being statistically more profound and less distorted by noise. The sensitivity of a ToF can be tuned via the pushing rate. The less pushes are performed per second, the more ions are measured per push. The mass resolution depends on the speed of the detection and timing system which has to be capable of detecting time-of-flight differences in the picosecond range. (de Hoffmann and Stroobant, 2001; Smedsgaard *et al.*, 2006)

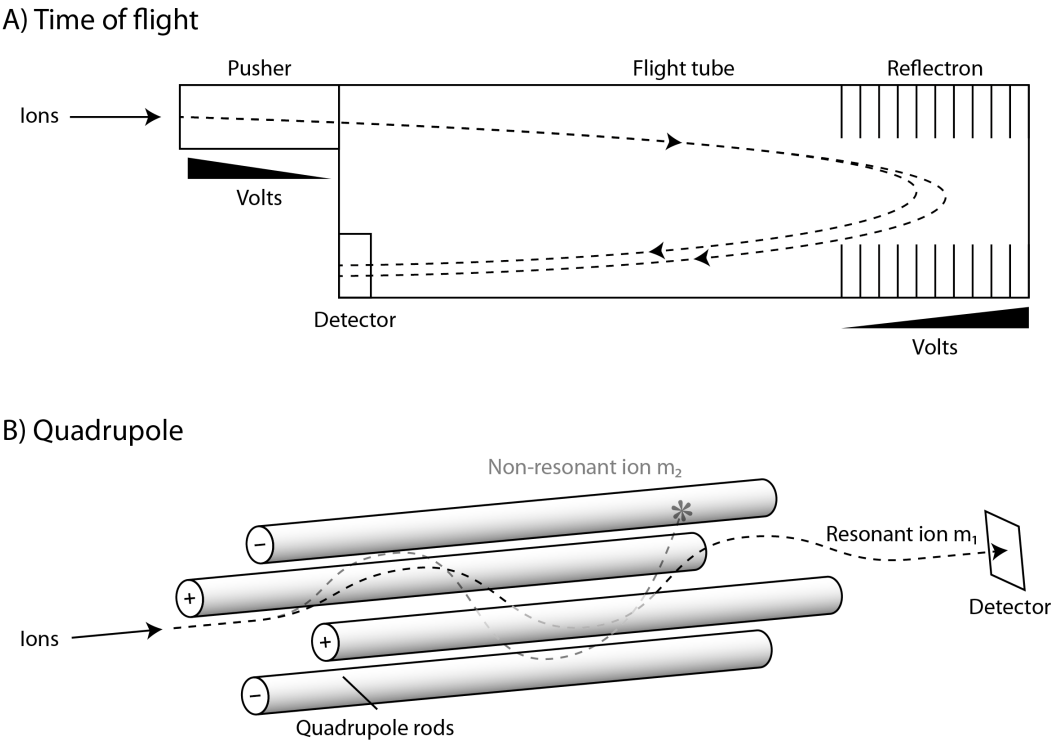

Figure 2.8: a) Schematic presentation of a reflectron time-of-flight (ToF) mass analyzer. Ions coming from the ion source are accelerated by the pusher before entering the flight tube. All ions carrying the same charge will have (about) the same kinetic energy but a velocity that depends on their m/z ratios. The flight tube is field-free except for the reflectron, which corrects variances in the kinetic energies of ions with the m/z values and thus increases the mass resolution of the instrument. The flight time t of an ion until it reaches the detector is proportional to the square root of its m/z ratio. b) Schematic presentation of a quadrupole mass analyzer. The quadrupole consists of four parallel cylindrical metal rods, of which the opposing ones are connected electrically to each other. A radio frequency (RF) voltage with a direct current (DC) offset is applied between the pairs of rods. The result is an oscillating electric field in which passing ions will have trajectories specific to their m/z ratios. Most of these trajectories are unstable and only ions with a specific m/z value, which depends on the parameterization of the quadrupole, will reach the detector. (de Hoffmann and Stroobant, 2001; Smedsgaard *et al.*, 2006)

Quadrupole mass analyzer

Quadrupole mass analyzers force ions into cylindrical trajectories, which mainly depend on an ion's mass and the direct current (DC) and radio frequency (RF) voltages applied to the quadrupole. This is explained in Fig. 2.8 b). If only a RF voltage is applied, the quadrupole serves as a wide pass filter, allowing ions of a wide range of m/z values to pass with a comparably high transmission. In two-dimensional mass spectrometry (MS/MS) these quadrupoles are commonly used to guide and focus ion beams, or - when filled with a collision gas like nitrogen, argon, or rarely helium - to induce a second fragmentation of ions colliding with gas molecules.

If a DC voltage is applied on top of the RF voltage, the range of m/z values for which ions are transmitted is narrowed down: The device becomes more selective and a mass separation is achieved, exceeding unit resolution if two $(m/z)_1$ and $(m/z)_2$ are more than 1 Da apart.

When continuously changing the voltages for DC and RF, the quadrupole allows to *scan* through m/z values in a set range, like e.g. from 100 m/z to 600 m/z, measuring a single (nominal) m/z at a time. In this case however, the instruments sensitivity is effectively reduced by the factor 500: From 1 second of elutes (from a GC and EI for example) ions of each m/z will only be transmitted for 2 ms. In targeted analyses the scanned m/z range should thus be narrowed down as much as possible; but this comes to the cost of informative fragmentation spectra ('diagnostic ions') which are necessary for the identification of unknowns (see section 2.5.4) (de Hoffmann and Stroobant, 2001; Smedsgaard *et al.*, 2006).

2.4 Preprocessing of chromatography-hyphenated MS data

The raw data, as it is written by the digitizer of the instrument, has to undergo a number of general preprocessing steps before it can be used for application-specific data integration and interpretation. These steps include mass calibration, baseline correction, noise filtering, sometimes chromatogram alignment and finally peak detection, normalization, and quantitation. This section is intended to describe the structure of chromatography-hyphenated MS raw data and to outline common preprocessing steps.

2.4.1 Structure of chromatography-hyphenated MS data

Generally, data from any MS coupled to any chromatographic method are a list of mass spectra s_t, where each spectrum is a set of pairs of mass-to-charge ratio and

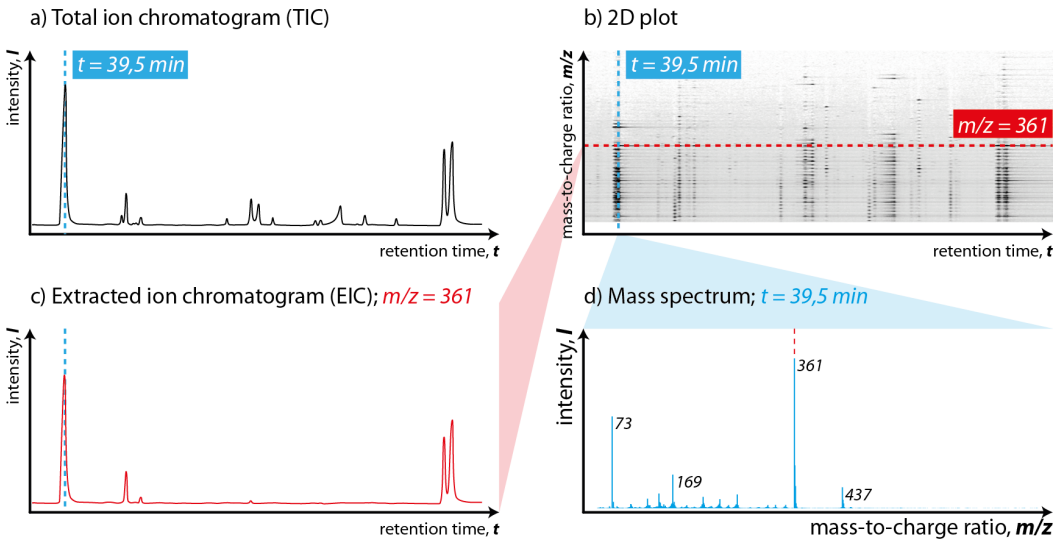

Figure 2.9: The structure of chromatography-hyphenated MS data, depicted using exemplary data. a) The total ion chromatogram (TIC) plots the summed-up intensity I of all ions measured to any retention time t. It is thus agnostic of MS information; b) The 2D plot additionally includes the MS information. The intensity is coded with grey values at retention time and m/z coordinates (darker grey value means higher intensity); c) The extracted ion chromatogram (EIC) plots the measured intensity at each retention time for ions of a certain m/z value and can be seen as a horizontal slice through the 2D plot or as a subset of the TIC; d) The mass spectrum plots intensities for all m/z values at one specific retention time and can be seen as a vertical slice through the 2D plot.

intensity $[(m/z)_{t,i}, I_{t,i}]$, acquired over a time dimension t (retention time). Fig. 2.9 summarizes the layers of this data structure and their interconnections.

The file formats in which these data is stored varies widely between vendors, but the metabolomics open-source and open-data community has agreed on certain file formats of which the flexible mzML format is currently recommended (Rocca-Serra *et al.*, 2016). Predecessors are the mzData and mzXML formats which have heavily influenced the development of mzML (Orchard *et al.*, 2007). Earlier the netCDF format has been *de facto* standard (Arita, 2004; Matthews and Miller, 2000). In practice all of these are still in use.

2.4.2 Mass spectral data preprocessing steps

Not necessarily but typically a number of preprocessing steps, concerning the MS derived data, are performed before writing to exchange formats. Often such preprocessed files are still referred to as *raw data*. These preprocessing steps are outlined in the following.

Filtering

At first the signal-to-noise ratio of spectra has to be improved in a filtering step. In the simplest case this is done using a moving average filter, removing high-frequency spikes from the profile spectra. The softening effect of the moving average window is controlled via its window size: The larger the window, the more spikes will be eliminated, eventually including valid peaks of interest. Unfortunately the moving average does not well preserve the original peak intensity and shape. The preferred option is to use a Savitzky-Golay filter (Savitzky and Golay, 1964), fitting a polynomial of order d to the moving window and evaluating it in its center point $(m/z)_c$ to replace its intensity $i_{(m/z)_c}$ with the filtered intensity $\hat{\imath}_{(m/z)_c}$ calculated from the polynomial. The filter can be further improved when including weighting functions in the estimation of the best order d.

Centroiding

Next, spectral data is usually centroided. I.e. most of the information in a profile spectrum is discarded in order to only preserve one representative data point for each ion peak. To achieve this, at first peaks are detected with help of an expected peak shape. Then each centroid is positioned to the center m/z of the top 50 % of the peak and the respective intensity (Hansen *et al.*, 2006). Centroiding is a massive data reduction step, hopefully discarding only data that is of no more use after this point in the data analysis.

Internal mass scale calibration

To remove systematic errors in mass accuracy, one or multiple internal mass references are injected into each analysis. The theoretical ion masses m_{lock_i} for each of these reference compounds $i = 1, ..., N$ must be known and are referred to as lock masses. Any measured m/z values matching one of the lock masses within a small tolerance window Δm will be used as anchor points to correct the mass scale, moving the anchor points to their exact (theoretical) mass values (Hansen *et al.*, 2006).

2.4.3 Chromatogram alignment

The retention times of all elutes from the chromatographic column are not perfectly reproducible across replicate measurements, but variations are introduced by small differences in pressure, temperature, solvent composition, column aging, and fluctuations (Hansen *et al.*, 2006). Retention time deviations vary from a few seconds in small batches to half a minute or more, if measurements have been run months or years apart. In analytical workflows that rely on compound identification per spectral similarity, these retention time shifts can be easily dealt with applying an appropriate threshold for maximum retention time deviations. However, if statistical analyses are to be performed on basis of single peaks, rather than compound spectra, retention time shifts can quickly introduce severe errors and misinterpretations to statistical results. In these cases it may be important to align all chromatograms in advance to further data analysis.

The aim of retention time alignment is to transform the retention times $t(p_k(c_i), s_j)$ of all peaks $p_{k\in L}(c_i)$ that derive from all compounds $c_{i\in N}$ found in all samples $s_{j\in M}$ to yield sample independent retention times $t^*(p_k(c_i))$. Such transformations are typically based on anchor peaks $p_{a\in L}$, either selected manually or determined automatically due to high signal intensities and high recovery rates among samples. Two samples s_i and s_{i+1} can then be aligned to each other by moving the retention times $t(p_a, s_i)$ and $t(p_a, s_{i+1})$ to the same $t^*(p_a)$. Even though symmetric retention time alignment methods exist, most of them are asymmetric - i.e. one sample is selected as a reference, and the retention times of anchor peaks in all other samples are moved to match this reference. Computing symmetric retention time alignments is resource intensive, as data from all chromatograms has to be compared to each other. In any case the retention times of peaks situated between anchor peaks are interpolated afterwards, e.g. linearly.

2.4.4 Peak detection and quantitation

Making the signals from the raw data accessible to quantitation and statistical analysis means to detect peaks in the data and to describe them with a position (RT and m/z) and a peak intensity and/or area. For this, peaks must be detected in both domains of the raw data, the m/z domain and the chromatographic domain. For both literature describes different approaches which are suitable for data from mass spectrometers with different resolutions (Danielsson *et al.*, 2002; Katajamaa and Oresic, 2005; Smith *et al.*, 2006; Tautenhahn *et al.*, 2008; Castillo *et al.*, 2011).

Lower resolution data is often binned in the m/z domain. That means the data is split into slices of e.g. 0.1 m/z width, making the slice available for subsequent processing in the chromatographic domain. These slices are sometimes referred to as extracted ion base peak chromatograms (EIBPC). The representative m/z value

for a peak in the EIBPC is later read from the raw data, for example by taking the value with the highest intensity. The disadvantages of this approach are mainly two-fold: First, analyte signals of a certain mass width may be split into two adjacent slices which results in 'jagged' EIBPCs. This issue can be mitigated by combining adjacent EIBPCs. The second, may be more severe problem is when the signals of two analytes fall into the same EIBPCs (Smith *et al.*, 2006).

The latter is the main reason why m/z domain binning is not a good option for high resolution data. Here it is favorable to do peak-centroiding (see 2.4.2) and to consider controids in the chromatographic domain if they can be found in at least n subsequent scans within a narrow m/z window Δ_{mz}, which is basically determined by the instruments mass resolution (Tautenhahn *et al.*, 2008).

Either way, the next step is to find peaks in the chromatographic domain. Chromatographic peaks are expected to have a close-to Gaussian shape, as has been explained in subchapter 2.3.2 Basics of chromatography and especially in Fig. 2.3. However, peak shapes are obscured by noise, which especially hampers detection of very low abundant signals. The most prominent approach to find and assess chromatographic peaks, i.e. to detect the beginning and the end of such a peak in the chromatographic domain, is matched filtration. Matched filtration with a second-derivative Gaussian wavelet overshoots the actual intensity of the peak, but its zero-crossing points mark the left and right borders of the peak. The area of the peak can now be assessed by integrating the raw chromatographic peak intensities between these left and right borders. The intensity can be assessed by either taking the local maximum or by reading the intensity off the center of the peak (Smith *et al.*, 2006).

The here described matched filtration with a second-derivative Gaussian function requires to parameterize the width of that function via its standard deviation σ, which should reflect the expected duration w_b it takes for an analyte to elute from the applied chromatographic column (see Fig. 2.3A).

It is now easy to remove background noise by simply discarding all peaks that do not exceed a certain signal-to-noise ratio S/N, where S is the intensity of the peak and the noise N can e.g. be estimated via the mean of all intensities in the chromatogram (Smith *et al.*, 2006). Sometimes the determination of S is also constrained to more local retention time windows. The S/N parameter, typically set in between $S/N = 2$ and $S/N = 10$, also allows to control the trade-off between sensitivity and selectivity of the peak detection.

2.5 Integration of metabolomics chromatographic data

After the acquisition and preprocessing of the data all the recorded information has to be gathered and brought into context with each other - within one measure-

ment, but also across analyses. This includes the allocation of all signals which derive from the same molecule, but also the exploitation of the measured data and its comparison with existing data from databases in order to annotate (or ultimately identify) the detected molecules. These measures for data integration are explained in the next subsections.

2.5.1 Spectra deconvolution

Both GC-MS and LC-MS create spectra that can consist of more than one mass signal (i.e. m/z value) per originally measured metabolite and get additionally polluted by background noise. In case of GC-MS spectra the multitude of mass signals is mostly due to the strong fragmentation in the EI and additional ions that come from prior derivatization to mask polar groups (see subsection 2.3.4). These GC-MS spectra are very reproducible and characteristic for their analytes. Together with the molecules' retention time they are often sufficient for compound annotation via database queries.

However, low signal-to-noise ratios obscure these spectra and very often spectra of coeluting compounds overlap (Halket *et al.*, 1999; Stein, 1999; Fiehn, 2002). Then, only mass signals that derive from the same metabolite need to be extracted from the raw spectrum to create a new, purified spectrum. This step is called spectra deconvolution and the artifically formed spectra are called pseudo spectra. Stein (1999) developed the first method for GC-MS spectra deconvolution. The method makes use of a model peak shape and the least-squares method to distinguish true features from noise and to deconvolute overlapping spectra.

For LC-ESI-MS data it is not sufficient to create purified pseudo spectra, because there are virtually no reference databases as spectra are far less reproducible here. Mass spectra created from LC-ESI-MS are different and pose their own unique challenges on the deconvolution step. During analysis with ESI, so called pseudo-molecular ions are created that can be observed as m/z values after detection. Pseudo-molecular ions are intact analytes (with a monoisotopic mass M) that build charged adducts with small inorganic ionic species (like $[M+H]^+$, $[M+Na]^+$, and others). If the type of adduct is known, M can be determined easily and may be subjected to mass decomposition or simple database queries.

Usually it is not that easy to get the monoisotopic mass M of the original metabolite though: The set of m/z values that can be found for each original metabolite is further complicated by fragmentation events that occur during ESI. Here, pseudo-molecular ions loose neutral groups (neutral losses) and the remaining fragments will be found as signal peaks with lower masses. Typical fragments for positive mode LC-ESI-MS are e.g. $[M+H-H_2O]^+$ or $[M+H-NH_3]^+$. It is thus mandatory to identify the roles of each peak in a pseudo spectrum to be able to calculate M.

Furthermore it is important to consistently determine the 'main peak' of a pseudo

spectrum, e.g. the $[M+H]^+$, to compare its intensity or area across samples in quantitative analyses.

The allocation of pseudo spectra is additionally hampered by a large amount of noise. Keller *et al.* (2008) have estimated that as little as 10 % of mass signals in LC-ESI-MS data are of true biological origin.

2.5.2 Spectra matching

To reidentify known metabolites in newly acquired samples, their mass spectra and retention times can be matched against databases of previously identified metabolites. Retention times are more reproducible in a GC context, rather than in an LC context, where chromatography methods differ widely. This subsection is dedicated to the matching of query spectra against reference spectra in databases. However, Bristow *et al.* (2004) showed that the reproducibility of mass spectra across different instruments is limited as well. It can thus be difficult to correctly assess the similarity of two spectra and different methods have been discussed in literature.

The simplest way of comparing two spectra is counting the peaks (m/z values) that are present in both spectra. More elaborate methods like normalized euclidean distances, absolute value distance, similarity index according to Hertz *et al.* (1971), probability-based matching, and the dot-product have been comprised by Stein and Scott (1994). Of them, the dot-product was rated best by Scheubert *et al.* (2013). In the context of this thesis spectra similarities are assessed using the cosine score, which is simply the square root of the dot-product. Both are explained in detail in the following.

To calculate the similarity of two spectra using the cosine score, it is at first important to create a vector of intensities for each spectrum. These vectors shall contain intensities for the same characteristic m/z values in the same order. There are three possible strategies to do this: Consider only m/z values that are present in both spectra, consider only m/z values that are present in the reference spectrum, or consider only m/z values that are present in the reference spectrum. A certain m/z error ε_{mz} must be allowed, to find equal m/z values in both spectra. Once the two intensity vectors have been created, it is possible to calculate the cosine between these vectors, as is listed in equation 2.9. The dot-product is simply the square of the cosine.

$$\cos \sphericalangle(\overrightarrow{Q}, \overrightarrow{R}) = (\overrightarrow{Q} \cdot \overrightarrow{R}) \, (|\overrightarrow{Q}||\overrightarrow{R}|)^{-1} \tag{2.9}$$

$\overrightarrow{Q}$ the intensity vector of the query spectrum.
$\overrightarrow{R}$ the intensity vector of the reference spectrum.

The cosine score for any two perfectly matching spectra is exactly 1. The advantage of this score is that if all intensity ratios within the query spectrum equal all intensity ratios with the reference spectrum, i.e. $\overrightarrow{Q} = \overrightarrow{R} \cdot x$, where x is any factor, the cosine is still 1. Furthermore, the absence of peaks in one spectrum that have a relatively low intensity in the other spectrum add only a small penalty to the score. This is very important whenever low intensity query spectra are matched, and some peaks potentially fell below the detection limit. If intensity ratios between the query and the reference spectrum differ too much, the penalty to the score is relatively high. Fig. 2.10 depicts four different query spectra that match the same reference spectrum and the respective cosine scores for each match.

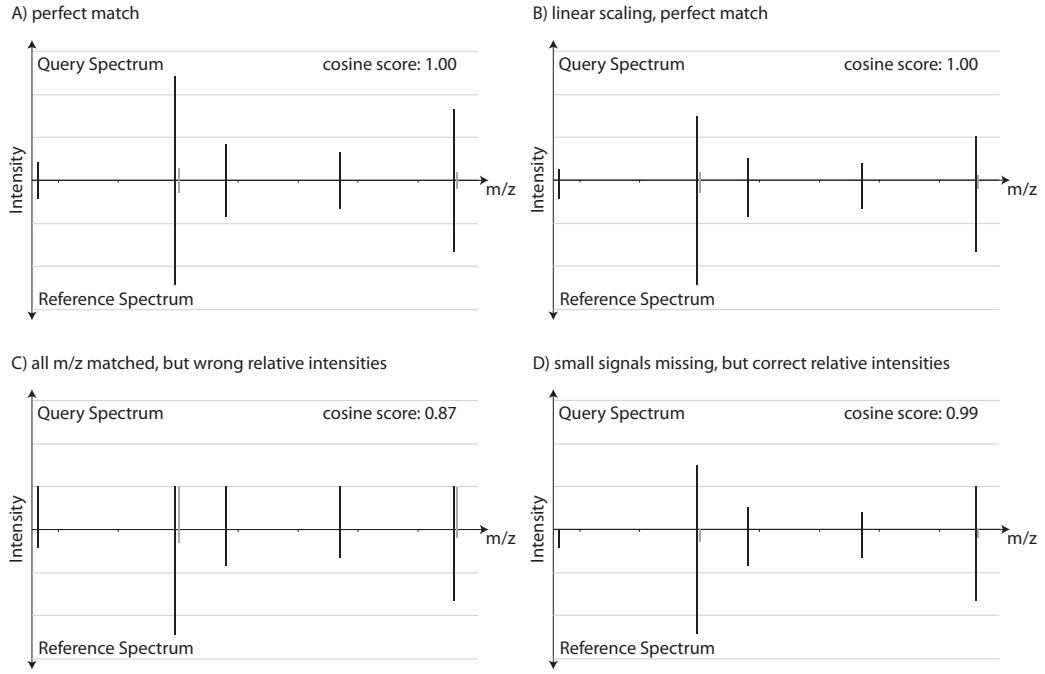

Figure 2.10: Four different query spectra matched against a reference spectrum. The cosine score has been calculated considering all peaks that are present in the reference spectrum. Missing intensities in the query spectrum have been set to zero. Isotope peaks are not considered. A) Perfecting matching spectra get a cosine score of 1; B) A linear scaling of all intensities in a vector does not affect the cosine score; C) If the intensity ratios of two spectra do not match, the penalty to the cosine score is quite high; D) Small missing peaks do hardly affect the cosine score, if other intensity ratios match.

After all, mass spectra are not yet unique and characteristic enough, to unambiguously identify (or reidentify) metabolites on a spectral basis alone. Spectra are often rather specific to compound classes, than to single compounds. In both GC-MS and LC-MS context it comes most natural to include the retention time into the decision process. Still, a true confidence cannot be assessed from database matches. Unlike for proteomics, it is not yet possible to establish decoy databases for small molecules, and thus False Discovery Rates (FDR) cannot be estimated. It still remains to the user to decide, how much confidence can be put in any database hit (Scheubert *et al.*, 2013). A more profound disucssion of the different levels of 'identification' can be found in subsection 2.5.4.

2.5.3 Mass decomposition

Metabolites cannot be sequenced like polynucleotides or polypeptides. The genome and the proteome are basically made of linear 4-letter and 21-letter codes, respectively (let aside post-translational modifications). Metabolites however are three dimensional arrangements of atoms. To unambiguously identify any unknown metabolite it must undergo structural elucidation. Unfortunately, MS technology alone is insufficient for complete structural elucidation of unknown compounds (Scheubert *et al.*, 2013). It is capable of elucidating their elemental composition though. Molecular formulas are very valuable annotations for metabolites and in part they can be derived from a mass spectrum.

The exact monoisotopic mass of a molecule is the basis to calculate its molecular formula, which summarizes the atoms that contribute to this molecule. In fact, every molecular formula has a unique monoisotopic mass, or vice versa: for a given monoisotopic mass only one combination of atoms fits.

For example the molecule S-adenosyl-L-methionine has a monoisotopic mass of 398.137238 Da which can only be explained by the molecular formula $C_{15}H_{22}N_6O_5S$ as is explained in equation (2.10).

$$\begin{aligned} 398.137238\ Da &= 15 \cdot m_{^{12}C} + 22 \cdot m_{^{1}H} + 6 \cdot m_{^{14}N} + 5 \cdot m_{^{16}O} + m_{^{32}S} \\ &= 15 \cdot 12\ Da + 22 \cdot 1.007825\ Da + 6 \cdot 14.003074\ Da \\ &\quad +5 \cdot 15.994915\ Da + 31.972070\ Da \end{aligned} \tag{2.10}$$

$m_{...}$ monoisotopic masses taken from the IUPAC Technical Report (de Laeter *et al.*, 2003), first six digits only.

However, no existing mass analyzer is capable of precisely determining the exact mass of any molecule. Modern instruments get down to a mass accuracy of 3

ppm, some even 1 ppm and below. For every day measurements errors of around 5 ppm should be assumed though. That is a range of 5 mDa for a molecule of 1,000 Da, which is not exact enough to uniquely identify the corresponding molecular formula. Thus, all possible molecular formulas must be calculated that fit a mass in the range of 999.995 Da to 1,000.005 Da.

Considering all known elements and any possible combination of these that fit $1,000 \pm 0.005\ Da$ would result in a vast and meaningless list of mostly arbitrary molecular formulas. In real world applications though, and certainly in a biological context, the atomic alphabet can be restricted to a small number of elements that are included in the mass decomposition. Still, the number of combinations growth exponentially with higher target masses and wider mass accuracies.

In metabolomics it is often sufficient to restrict the elemental alphabet for mass decomposition to C, H, N, O, P, and S. Still, for masses in the range of $1,000 \pm 0.005\ Da$ more than 30,000 formulas can be found - more than 300,000 if the mass accuracy is as low as $0.05\ Da$. Fig. 2.11 depicts how the number of theoretical mass decompositions depends on mass and mass accuracy. But most of the resulting formulas are very improbable to exist or even are chemically impossible. Kind and Fiehn (2007) have comprised a set of rules that can be applied to filter mass decomposition results. This set includes chemical rules considering atom connectivity (Senior, 1951) or isotopic patterns (Kind and Fiehn, 2006) as well as heuristic rules that were evaluated using databases of biological compounds.

In the case of LC-ESI-MS measurements additional information may be derived from spectra deconvolution (see section 2.5.1). Neutral losses, if successfully identified, reveal sub formulas of the true elemental composition: molecular formulas can be excluded from the possible results, if they do not contain the molecular formula of the neutral loss (Rojas-Chertó *et al.*, 2011; Kessler *et al.*, 2014).

2.5.4 Metabolite identification

The term 'metabolite identification' is multilayered. First, it has to be differentiated between metabolite identification in the sense of annotating a signal peak or spectrum with the name of a known compound, and the so-called *de novo* identification of a hitherto unknown molecule. Second, the criteria that are required to identify a metabolite with reasonable confidence, with respect to precedented or *de novo* identification, is a matter of discussion and community efforts for more than a decade now. This section will summarize the currently prevailing proposals on different qualities and confidences of metabolite identification. As *de novo* identification cannot be done with MS and chromatographic methods alone, but needs to be supported by orthogonal technologies like NMR (Bino *et al.*, 2004; Scheubert *et al.*, 2013), details on this topic are beyond the scope of this thesis and will not be covered in this section.

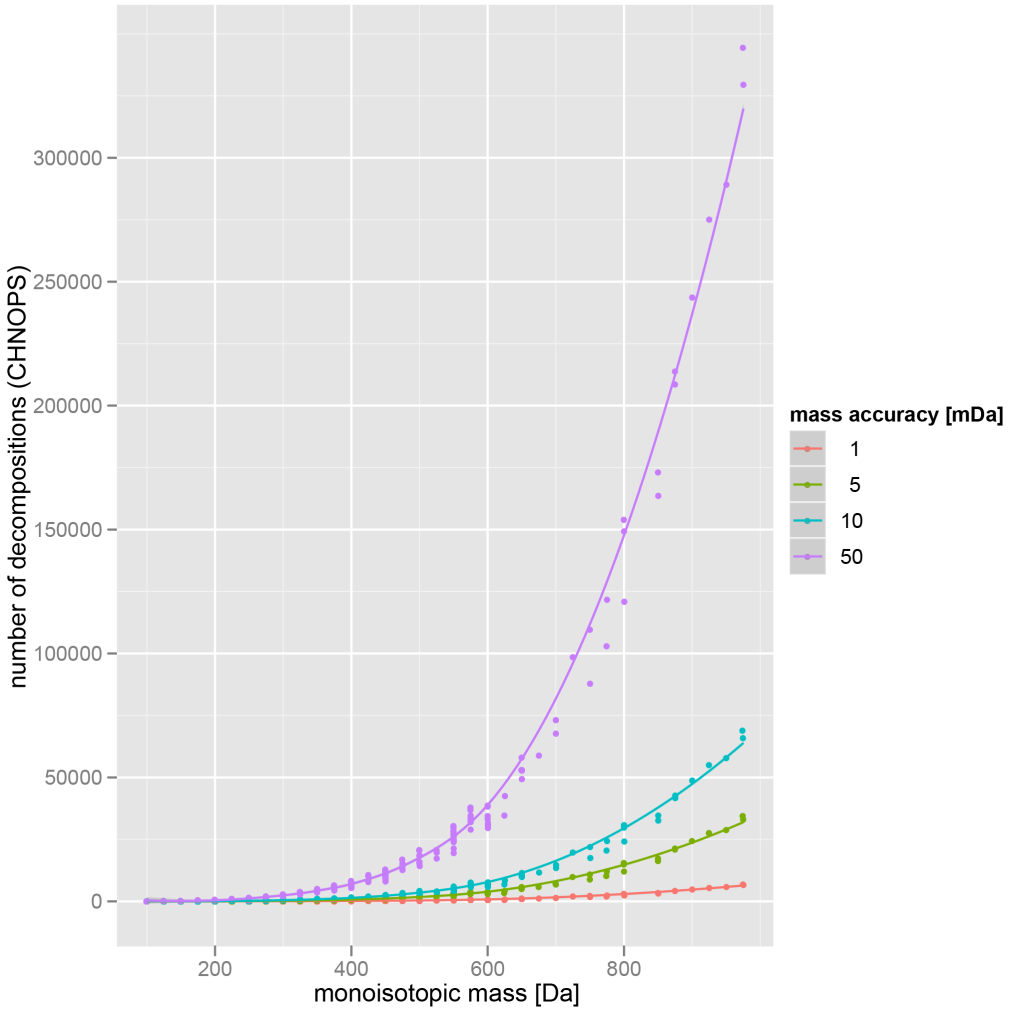

Figure 2.11: The number of calculated mass decompositions in dependency of the target monoisotopic mass and the mass accuracy for the atomic alphabet of $L = \{C, H, N, O, P, S\}$. For nominal masses from 100 Da to 1000 Da, in 25 Da steps, compounds from the KEGG COMPOUND database have been selected. For each of these compounds monoisotopic masses, decomposition with four different mass accuracies was performed (1 mDa, 5 mDa, 10 mDa, and 50 mDa) using the mass decomposition algorithm of the ALLocator software.

In 2007 Fiehn *et al.* published a brief report announcing the metabolomics standards initiative (MSI) to develop a set of reporting standards for metabolomics in order to catch up with the high standards that were already established in the proteomics research community. In the same issue of *Metabolomics* Sumner *et al.*, the MSI chemical analysis work group (CAWG), presented a first primer suggesting to consider four different levels of metabolite identification: compare Tab. 2.3.

Table 2.3: Four levels of metabolite identification confidence as proposed by the chemical analysis work group of the metabolomics standards initiative (Sumner *et al.*, 2007).

Level	Name and Description
1	Identified compounds The compound has been unequivocally identified by comparing two or more orthogonal properties of an authentic standard reference to the experimental data. Reference and experimental data must have been analyzed with the same methods in the same laboratory.
2	Putatively annotated compound Compounds that were identified based on physicochemical properties and/or spectral similarity to known compounds (see section 2.5.2 on spectra matching.)
3	Putatively characterized compound class The compound class could be identified based on physicochemical properties and/or spectral similarity to other, known compounds of that chemical class.
4	Unknown compounds Neither the compound, nor its chemical class could be identified. But based on its spectral data it can still be differentiated and quantified.

Even though the reporting standards recommended by the MSI were widely regarded as the 'community consensus' for seven years, they have hardly been applied. Neither scientific journals nor public metabolomics databases added authority to the standards by making them a requirement for publications (Salek *et al.*, 2013b). The four levels have been a simple system, but they did not sufficiently differentiate the levels of confidence, which can be found as subsets of the levels 2 and 3. Furthermore, the four level system was technology-agnostic and as such not expressive enough for reporting metabolite identification confidence

(Creek *et al.*, 2014).

In 2014 then, Schymanski *et al.* and Sumner *et al.* reacted to the request to solve the issues of the four level standard and suggested more detailed systems to assess identification confidence.

Schymanski *et al.* proposed a new five level system, tailored to high resolution MS, that not only covers identification confidence, but also better integrates the accuracy of the identification: Is exactly one candidate left, can the molecular formula be determined or only the exact mass (cmp. 2.4)?

Table 2.4: Five levels of metabolite identification confidence in high resolution mass spectrometry as proposed by Schymanski *et al.* (2014).

Level	**Name and Description**	**Min. data requirements**
1	Confirmed structure by reference standard	MS, MS/MS, RT, Ref. Standard
2	Probable structure	
	a) by library spectrum match	MS, MS/MS, Library MS/MS
	b) by diagnostic evidence	MS, MS/MS, Experimental Data
3	Tentative candidate(s) structure, substituent, class	MS, MS/MS, Experimental data
4	Unequivocal molecular formula	MS isotope/adduct
5	Exact mass of interest	MS

However, this five level system still lacks information on additionally applied technologies (e.g. orthogonal measurements like NMR) and is still not much more fine-grained than the four level system by MSI. These are the shortcomings that Sumner *et al.* (2014) tried to tackle with their proposal of an actually quantitative metric and an alphanumeric code for metabolite identification confidence. For example an LC-ESI-MS measurement yielding an exact mass with an accuracy below 5 pmm, would be worth 1 point. A high resolution retention time would be worth another 1.5 points. The sum would be 2.5 points and might be multiplied by 2, when matching to a (public) reference library. This would result in a final score of 5 points. The points that each technology 'is worth' were carefully selected by Sumner *et al.*, but may only be seen as a primer for future discussion. Not quantitative but more definite is the alphanumeric code that is suggested in the same publication. The above example would be encoded as $HRMS^{1}_{PL}, HR_{tPL}$,

where $HRMS^1$ indicates a 1-dimensional high resolution mass spectrum and HR_t notes the high resolution retention time. The subscript PL states a match against a public library.

The five level system by Schymanski *et al.* appears to be easier and more generalized, since levels 3 to 5 allow to describe annotations that comprise a set of compounds, like substituents, compounds with a common molecular formula or monoisotopic mass. The quantitative and alphanumeric code by Sumner *et al.* better reflects the technologies and accuracies that led to an annotation. None of the yet proposed systems for reporting metabolite identification confidence takes into account, how unequivocal molecular formulas were determined from the monoisotopic mass: e.g. interpretation of isotope patterns and/or fragmentation patterns are not considered.

2.5.5 Preparation of quantitation tables

In almost all cases it is required to perform some kind of statistics to finally generate new information off a metabolomics experiment. In a first step, the measured data is thus brought into a matrix X of n samples and m variables. Variables will be mostly constituted of measured features (signals, isotope clusters, or compounds), but may also contain 'external data' which was not acquired with an MS instrument, like e.g. age, nutrition, temperature etc. A challenge in this step, which may not be underestimated, is the task of associating features across the samples, or in other words, to concatenate the feature vectors $n_i = [x_0...x_m]$ of all samples to a single matrix. Signals vary in their positions because of retention time shifts and mass errors. These problems can be tackled with one of multiple strategies, like binning, retention time correction, or matching on spectral basis, which have been discussed above. Another issue is that not all signals that are present in one sample, are present in the next sample too - typically not even in technical replicates. These signals that are absent in only a subset of samples are referred to as missing values.

Missing values

These missing values can be results of noise in the other samples (i.e. the *present* signals are artifacts), of the complete absence of the compound, of an abundance below the detection limit, or as a result from misses during preprocessing or too strictly parameterized data integration. It is thus not necessarily correct to account for them with zero intensity or area. Also, certain statistics cannot deal with missing values due to division-by-zero problems. There is a multitude of strategies on how to deal with missing values. Values may be set to the minimum value found for the same feature, or to an arbitrary low value that should be as close as possi-

ble to the detection limit. These two strategies are helpful, if signals were lost due to low intensities. Values may alternatively be replaced by the mean or median of either all findings, or only derived from findings within the same group. The latter strategies are helpful, if the signals are expected to be present in higher intensities, but supposedly were missed or falsely assigned during data processing.

2.6 Analytical approaches to the metabolome

The analysis of one or many metabolites in a sample may be motivated by a number of generally different targets. The priority may be to find a certain metabolite and to measure its relative abundance, or it may be important to have an unbiased overview of as many metabolite levels in a complex sample as possible, or may be single metabolites are not important at all, but unelucidated mass spectra are used to classify an organism. The different takes on metabolite analysis, with decreasingly focused scopes but increasing coverage of the metabolome, can be referred to as *targeted analysis*, *metabolic profiling*, *untargeted analysis* (sometimes called *metabolomics*), and *fingerprinting*. These different approaches are presented in this section. Orthogonally to these approaches, *stable isotope labeling* (SIL) allows for relative quantitation and facilitates the annotation of molecules. SIL is outlined in the last subsection.

2.6.1 Targeted analyses

If finding and quantifying a specific, well-known metabolite (or a few specific metabolites) is the goal, a *targeted analysis* is performed. The sample preparation is tailored to recover the metabolite of interest while washing off as many other metabolites as possible. Whether applying GC-MS or LC-MS, the RT as well as the m/z values of characteristic ions are known in advance. It is thus easy to find the respective signals in all measurements. Here the sensitivity and the dynamic range of the entire analytical approach (including sample preparation, chromatography and mass spectrometry) are key, so that very low abundant molecules can be found and that abundances can be assessed within a high dynamic range.

As an example, *targeted analyses* can be applied to study the primary effect of the genetic alteration of an enzyme, where the analysis can be constrained to its substrate and/or product (Fiehn, 2002). As the list of target analytes growth, the border between *targeted analyses* and *metabolic profiling* blurs.

2.6.2 Metabolic profiling

Often *metabolic profiling* is not distinguished from *targeted analysis*, as differences are small. Here, the list of targets may e.g. be comprised of all metabolites in

a certain metabolic pathway. Typically, these metabolites share certain chemical characteristics, such that sample preparation can still be tailored to reduce matrix effects (Fiehn, 2002). As the metabolites are known in advance, the m/z values of expected ions can be deferred easily. Not necessarily all RT are known *a priori*.

Both *targeted analyses* and *metabolite profiling* are commonly driven by a specific biochemical question or in order to test a certain hypothesis (Patti *et al.*, 2012).

2.6.3 Untargeted analyses (metabolomics)

Untargeted analysis usually is an approach that is applied for hypothesis generation, rather than testing. There is no defined list of metabolites of interest before the experiment has been performed and the measurements have been evaluated. In contrast, often the list of potentially interesting metabolites is the result of the experiment. For example, when investigating the pleiotropic effects of a genetic alteration, multiple metabolites across various metabolic pathways may be affected (Fiehn, 2002; Weckwerth, 2003; Patti *et al.*, 2012).

Accordingly, untargeted analyses are designed to be as unbiased as possible, trying to include the entire metabolome. That means that sample preparation should avoid the exclusion of any class of metabolites and that the analytical method must provide sufficient resolving power to separately acquire the signals of a complex mixture. The resulting data is equally complex and adequate software tools are required: First to associate signals to each other if they derive from the same original molecules. And second to identify expected metabolites ('known knowns') as well as other metabolites described in literature and databases ('known unknowns'). In fact, data evaluation of untargeted analyses should also be prepared for the discovery of hitherto unknown metabolites and yield first hints for their *de-novo* identification.

LC-MS can cover more of the metabolome in a single measurement than GC-MS, but generally both technologies are suitable for untargeted analyses. From both it is also possible to derive RT, accurate mass and putative molecular formulas for yet unidentified molecules.

2.6.4 Fingerprinting

Assessing the metabolome with MS can even be helpful without identifying the metabolites that are responsible for each signal. The complex but very characteristic spectra themselves can be used for the classification of samples which is a powerful tool for clinical diagnostics. Such *fingerprinting* approaches do without any chromatography and can thus be performed rapidly, in a high throughput manner (Fiehn, 2002). Fingerprinting is mentioned here for the sake of completeness, but it is out of the scope of this thesis, which focuses on data evaluation for

chromatography-hyphenated MS.

2.6.5 Stable isotope labeling

Stable isotope labeling is not an alternative approach, but a possible extension for targeted as well as untargeted analyses. It is based on the replacement of the atoms of a specific element by a specific stable isotope in a certain culture or plant. For example all ^{12}C atoms (which in nature is the most abundant isotope of carbon) can be replaced by ^{13}C isotopes. That increases the monoisotopic mass $M_{i,C^{12}}$ of each metabolite by $\sim n_{C,i}\ Da$ to $M_{i,C^{13}}$, where $n_{C,i}$ is the number of carbon atoms in the molecular formula of M_i; for a hexose molecule $C_6H_{12}O_6$ that is a $\sim 6\ Da$ shift, accordingly.

This can be achieved by only providing ^{13}C carbon sources to an organism, which is way easier for bacterial cultures (e.g. grow with ^{13}C-Glucose), than for plants (prepare an atmosphere that contains $^{13}CO_2$ instead of $^{12}CO_2$) (Hegeman *et al.*, 2007; Giavalisco *et al.*, 2009). When the *in-vivo* labeling was successful, almost all ^{12}C moieties have been metabolized and about 99 % of all carbon atoms in the culture will be ^{13}C atoms. Fig. 2.12 explains how this affects the masses, and consequently the mass spectra, of metabolites.

There are two benefits of such experiments, when e.g. samples of a normally grown mutant strain and a U-^{13}C-labeled wild type strain are mixed and analyzed in a single LC-MS measurement. First, the U-^{13}C can serve as an internal standard: ^{12}C and ^{13}C peaks in a spectrum have to be normalized to the respective dry-weights, but then their quotient directly reveals the quantitative ratio. Second, the m/z distance between associated ^{12}C and ^{13}C peaks reveals the number of carbon atoms in the analyte. This information can be used to improve the identification of metabolites by reducing the number of possible molecular formulas (Rodgers *et al.*, 2000; Baran *et al.*, 2010; Rojas-Chertó *et al.*, 2011; Kessler *et al.*, 2014).

As carbon is a constituent of every metabolite, it is most often ^{13}C that is introduced to replace the naturally most abundant ^{12}C atoms, but ^{15}N and ^{34}S are suited as well, when investigating nitrogen- or sulfur-containing metabolites, respectively.

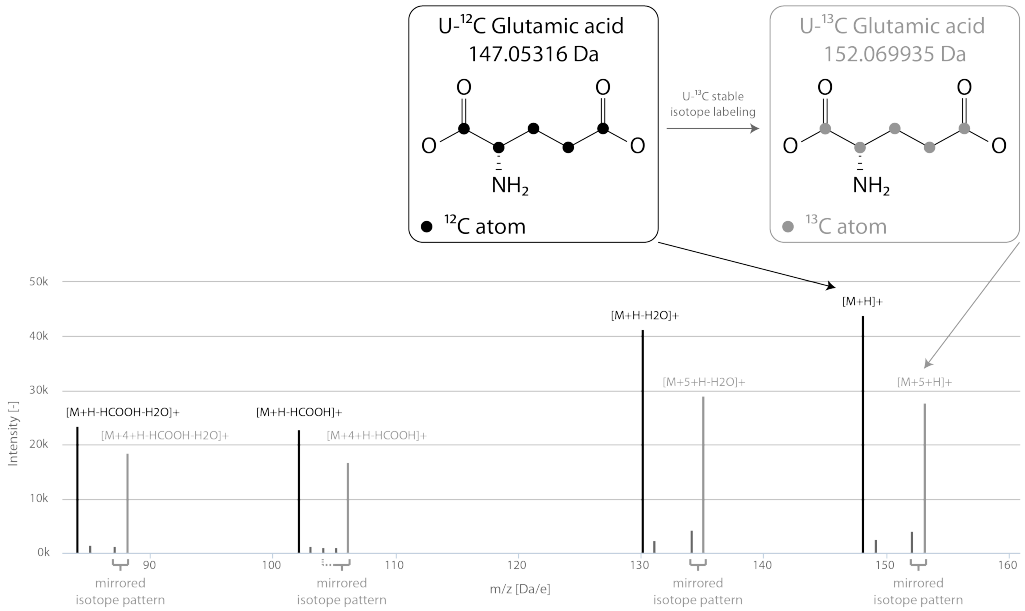

Figure 2.12: LC-ESI-MS spectrum of glutamic acid after deconvolution. The spectrum was taken from an ALLocator experiment to analyze complex samples of *C. glutamicum* measurements containing an internal standard with U-^{13}C SIL (Kessler *et al.*, 2014). While the figure was modified afterwards, no spectral information has been changed. Glutamic acid contains $n_C = 5$ carbon atoms. The U-^{13}C SIL consequently increases the mass of the molecule by $n_C \times 1.003355\ Da$, the mass difference between ^{13}C and ^{12}C isotopes, which sums up to $\Delta_m = 5.016775\ Da$. This mass difference can be observed for the $[M+H]^+$ and $[M+H-H_2O]^+$ ions. The other ions though, which have each lost one carbon atom in the ion source, both differ by only $4.01342\ Da$ accordingly. Some U-^{13}C glutamic acid molecules still contain one or more ^{12}C atoms, such that the respective isotope patterns feature signals for the lighter isotopologues. Such reversed isotope patterns are referred to as 'mirrored isotope patterns' in this thesis.

3 Current progress in metabolomics data analysis

Due to the diversity of instrumental set ups and experimental approaches that are applied for metabolome analysis, the metabolomics software landscape consists of many solutions that cover rather narrowly defined purposes. For instance, the fundamental differences in LC-MS and GC-MS spectra require different informatic approaches for interpretation and identification. For the same incompatibility reasons separate spectral data repositories are required. An example for experimentally motivated differing requirements to analytical software is between targeted and untargeted measurements.

More complex metabolomics software platforms and tool suites, which at least provide some parts of the metabolome data analysis workflow, are presented in the following sections. The applications are divided into command-line tools, mostly lacking graphical user interfaces and often focused on early processing steps, desktop applications, which typically excel through rapid raw data access, and web applications, which often profit from the compute power of their backing servers and sometimes take advantage of their potential as sharing and collaboration platforms. The last section comprises the most important public resources and repositories.

3.1 Command-line tools

This section is dedicated to freely available tools which do not (necessarily) feature a graphical user interface but that can perform one or a few steps of the metabolomics data processing workflow. They often aim to solve the early processing steps like peak detection, spectra deconvolution, or compound identification. Many of these tools are integrated in more comprehensive software platforms and thus constitute an important foundation for metabolomics software development.

3.1.1 XCMS

XCMS is an open source R package for peak detection, peak matching across samples, and retention time alignment in all types of chromatography-coupled MS,

but primarily LC-MS (Smith *et al.*, 2006). The original peak detection of XCMS applied matched filtration based on a second derivative Gaussian model peak shape to binned m/z slices in the raw data. In 2008 Tautenhahn *et al.* enhanced XCMS with the *centWave* algorithm, which detects regions of interest based on signal density (instead of binning) and a continuous wavelet transform based approach for peak detection (instead of the matched filtration) to overcome the bias of the predefined model peak shape. XCMS^2 is an extension for XCMS which supports matching for MS^2 spectra against the METLIN database. To perform second-order analyses on XCMS output of multiple groups the metaXCMS R-package was released (Tautenhahn *et al.*, 2010).

3.1.2 CAMERA

Another open source R package that builds upon XCMS output is the CAMERA tool for spectra deconvolution and annotation (Kuhl *et al.*, 2012). CAMERA groups peaks in XCMS peak lists and annotates them as adducts, fragments or isotopes deriving from putative molecules. By this, the accurate masses of these putative molecules are determined. CAMERA uses a graph-based approach and a predefined list of common adducts and fragments. It integrates seamlessly with XCMS.

3.1.3 MET-COFEA

Similarly to CAMERA, MET-COFEA (METabolite COmpound Feature Extraction and Annotation) aims to cluster chromatographic peak features by retention time and peak shape, and to subsequently annotate the most common adducts, fragments and their isotopes which reveals the molecular mass of the corresponding metabolite. MET-COFEA is applicable to LC-MS data and applies a mass trace based-extraction of ion chromatograms and a continuous wavelet transform-based peak detection (Zhang *et al.*, 2014). Identified peak groups can be used to align several ESI-based measurements.

3.1.4 X^{13}CMS

X^{13}CMS is an extension to XCMS which was described above. It allows to integrate ^{12}C and ^{13}C measurements (or any other isotope labels) and to use their feature ratios for differential analyses. X^{13}CMS builds upon the peaks and peak groups XCMS generates and detects the enriched features in them (Huang *et al.*, 2014).

3.1.5 AMDORAP

Takahashi *et al.* (2011) seek to get the most out of the accuracy advantage which LC-MS instruments have compared to other technologies. The open-source R

package AMDORAP was developed to obtain the most accurate *m/z* values from a set of measurements.

3.1.6 apLCMS

Similarly to AMDORAP, apLCMS is dedicated to peak detection in high-accuracy LC-MS data. This R package applies a set of adaptive procedures in order to avoid hard cutoffs (Yu *et al.*, 2009). In contrast to AMDORAP, the apLCMS software relies on a profile-by-profile analysis, rather than performing peak picking and alignment in a single step.

3.1.7 mzMatch(-ISO)

Introducing the file format PeakML, Scheltema *et al.* (2011) also presented a new R package mzMatch, which extends XCMS (Smith *et al.*, 2006) with functionality to read and write PeakML files. The PeakML format was designed to allow researchers to set up their data processing pipelines in a more modular fashion, rather than limiting themselves by using one or a few tools for the entire process. To overcome the limitations of static graphs, the PeakML Viewer desktop application was released, which allows to interactively explore the respective results. In 2013 Chokkathukalam *et al.* extended the mzMatch R software with the new package mzMatch-ISO. The latter aims to detect and visualize isotopes in SIL-experiments.

3.2 Desktop applications

In this thesis, a desktop application is defined as any software that must be installed on a computer, makes use of the local compute power, provides a graphical user interface, and persists raw data as well as results on this computer. Traditionally desktop applications provided the more sophisticated graphical user interfaces compared to web applications. However, this advantages is increasingly mitigated by the steep development web technologies have undergone in recent years. A remaining advantage is that desktop applications, which mostly interact with the local file system, are less prone to bandwidth limitations. This allows very detailed data representations to be updated in real-time.

3.2.1 mzMine

The Java-based open-source toolbox mzMine provides a collection of methods for the differential analysis of LC-MS data (Katajamaa and Oresic, 2005; Katajamaa *et al.*, 2006). It is designed to easily incorporate new algorithms and methods for

data processing, and to this end it was leveraged to an even more modular framework in its second version (Pluskal *et al.*, 2010). The mzMine software offers visual data exploration for quality assessment, but not for curation. Statistical analyses are not part of the mzMine workflow.

3.2.2 MetAlign

MetAlign, first published in 2009 by Lommen, is an interface-driven desktop application covering the analysis process of both GC-MS and LC-MS from raw data input, to preprocessing, to the export of spreadsheets for subsequent (external) statistical analysis. In its third version it is a multi-threaded application, which led to a performance leap in several processing steps (Lommen and Kools, 2012).

3.2.3 Decon2LS

Decon2LS is an open source software package (written on the .NET-framework) for automated processing and visualization of high-resolution LC-MS data. It can be applied for both proteomics and metabolomics data. Raw data reading, peak finding, modeling of theoretical isotope distributions, and deisotoping are covered. In-depth exploration of results is made possible with detailed visualizations. For compound identification purposes however, data must be exported and analyzed with other programs (Jaitly *et al.*, 2009).

3.2.4 PeakML Viewer

The PeakML Viewer (Scheltema *et al.*, 2011) has been described in conjunction with the PeakML file format and the mzMatch R package in subsection 3.1.7.

3.2.5 MetExtract

MetExtract is a desktop application (sources available online, precompiled for Windows operating systems) that allows to exploit SIL in a way that is very similar to ALLocator: It is mainly based on the mirrored isotopic patterns produced by e.g. 99 % ^{13}C labeled samples. The software uses the presence of both isotopic and mirrored isotopic patterns as a hard quality criterion and bases further analysis only on those features that have been found in both conditions (Bueschl *et al.*, 2012, 2013a,b). Important improvements that have recently been released with MetExtract II are metabolic feature pair detection and ion deconvolution (Bueschl *et al.*, 2017).

3.3 Web applications

Here, the term 'web application' comprises any software that can run in a users web browser, or is started from a users web browser (like Java™ Webstart applications). This covers a broad spectrum of programs from one-trick-ponies that perform a single analysis on uploaded data and then provide the results, to full-fledged software platforms including management and storage functionality. They all have in common that users do not have to install any additional software besides a web browser, and may be Java Webstart or the Adobe™ Flash™ Player. Consequently, software updates are centralized at the providers server and do not have to be distributed. Many web applications are also covered by the term 'Software as a Service' (SaaS).

3.3.1 MetaboAnalyst 2.0

The MetaboAnalyst web platform was first published in 2009 by Xia *et al.* and provided strictly defined solutions for the analysis of GC-MS, LC-MS and NMR data. With a second version especially the data analysis features were enhanced (Xia *et al.*, 2012). MetaboAnalyst makes use of the diverse software packages provided for the R project (R_Development_Core_Team, 2011) and Bioconductor (Gentleman *et al.*, 2004). With help of these, it covers the metabolomics analysis workflow from data preprocessing, to normalization, to statistical analyses, and finally generates analysis reports that can be downloaded.

3.3.2 XCMSOnline

XCMS (Smith *et al.*, 2006) is one of the most accepted open source tools for LC-MS peak detection in the scientific community. With XCMSOnline (Tautenhahn *et al.*, 2012) a graphical web interface was published that is completely dedicated to the tool. While not being the first online platform to allow XCMS usage (Neuweger *et al.*, 2008), XCMSOnline is tailored to provide the entire XCMS feature set. It is additionally enhanced by the features of CAMERA, which is another R package that seamlessly integrates with XCMS to perform spectra deconvolution (Kuhl *et al.*, 2012).

3.3.3 MetabolomeExpress

The MetabolomeExpress project aims to be a public repository for GC-MS data that also offers to (re-)perform analyses in that platform (Carroll *et al.*, 2010). To this end it provides its own peak detection algorithm as well as matching against public databases and a set of statistical tools to explore gained results.

3.3.4 MetiTree

The MetiTree web application was designed to support researchers in metabolite identification using multi stage mass spectrometry (MS^n) data (Rojas-Chertó *et al.*, 2012). The software provides means for data organization, processing, sharing, visualization, and matching, as well as a spectral tree viewer for exploration. MetiTree integrates the Multi-stage Elemental Formula (MEF) tool, which was developed by the same group, to assign elemental compositions to the ions and fragments in MS^n spectra (Rojas-Chertó *et al.*, 2011).

3.3.5 metaP-Server

Kastenmüller *et al.* (2011) published a web server for statistical analyses in metabolome research, including support for multiclass statistics. The metaP-Server does not offer preprocessing by itself, but it accepts quantitation data tables as input and is thus independent of the applied technology (MS- or NMR-based).

3.3.6 MetFrag

The combinatorial fragmenter MetFrag provides an interface and algorithm for metabolite identification with respect to their fragmentation patterns. MetFrag first queries public databases (HMDB, KEGG, or ChemSpider) by neutral exact mass or molecular formula to find candidate compound structures. Then for each compound the theoretical fragmentation pattern is determined and matched against the user's measured spectrum (Wolf *et al.*, 2010).

3.3.7 ChromA

ChromA is a web-based tool for the retention time alignment of chromatography-coupled MS data. It is applicable in both contexts, metabolomics and proteomics. For computing pairwise alignment, dynamic time warping is applied. In addition, the user interface allows to manually define additional anchor points - e.g. the positions of already identified compounds. Applying center-star approximation a reference file can be selected, to allow for alignment of more than two chromatograms (Hoffmann and Stoye, 2009).

3.4 Public resources and repositories for metabolome informatics

Regardless of the underlying platform all mass spectrometry software solutions that aim for annotation or even identification of small molecules need to match

spectra against databases. These may be user-created reference lists, but more often the tools query public repositories of compounds and spectra. Search criteria may be putative masses of compounds, complete or pseudo spectra, or even MS^n spectra. In 2013a, following the model of proteomics, even repositories for raw data of metabolomics experiments are created and populated (Salek *et al.*). The following section however comprises public databases that can be used by software tools to perform the task of spectra annotation (cmp. Fig. 4.1).

3.4.1 Kyoto Encyclopedia of Genes and Genomes

KEGG (Ogata *et al.*, 1999; Kanehisa and Goto, 2000) mainly consists of the three databases PATHWAY, GENES, and LIGAND. The latter comprises the three sections COMPOUND, ENZYME, and REACTION. In the scope of metabolomics the COMPOUND database is most interesting of course. But also in which reactions the molecules take part and in which pathways these are embedded is an important information that KEGG searches can return. The KEGG databases do not store spectra, but compounds can be filtered by their molecular masses.

3.4.2 Metlin

The Scripps Center For Metabolomics hosts Metlin, a freely available data repository for metabolite identification through mass analysis (Smith *et al.*, 2005). It allows for mass spectral searches and even stores MS/MS spectra for many molecules. Results are also linked to KEGG compounds where possible.

3.4.3 Human Metabolome Database

The HMDB is a comprehensive database of endogenous metabolites, comprising data from literature, experimental data, and MS as well as NMR spectra. For each available compound a so called MetaboCard is available, which contains more than 100 data fields with physico-chemical, biological, or biomedical data (Wishart *et al.*, 2007, 2009). The HMDB web platform offers mass spectral searches, but does not externalize the application programming interface (API).

3.4.4 National Institute of Standards and Technology

The National Institute of Standards and Technology (NIST) Standard Reference Database 1A (NIST/EPA/NIH, 2014) in its current version contains more than 270,000 EI spectra, almost as many MS/MS spectra, and Retention Index (RI) values for more than 80,000 compounds.

3.4.5 GOLM Metabolite Database

The GMD has been established as a metabolomics building block of the CSB.DB (Steinhauser *et al.*, 2004), a systems biology database project. The GMD allows to query a database of GC-MS spectra by spectral information and found RIs to obtain compound names and associated information (Kopka *et al.*, 2005).

3.4.6 ChemSpider

The ChemSpider database provides chemical, biochemical, and spectral information on more than 32 million structures from several hundreds of datasources in a single website. Its contents curation is based on a crowdsourcing strategy (Pence and Williams, 2010; Royal Society of Chemistry, 2014). It allows searches based on various query inputs, including identifier, structure, and monoisotopic masses to name just a few.

3.4.7 MassBank

MassBank is another public repository for mass spectra of compounds that are relevant to life sciences (Horai *et al.*, 2010). It contains more than 41,000 spectra obtained from different ionization methods according to the latest statistics from February 2015[1]. MassBank provides a SOAP API which allows to query the database from external software. With MoNA (MassBank of North America[2] a second facility takes on a central, auto-curating mass spectral database, currently storing beyond 200,000 spectra from more than 70,000 different compounds.

3.4.8 FiehnLib

The FiehnLib database contains EI-spectra and retention indices for more than 1,000 primary metabolites with masses below 550 Da (Kind *et al.*, 2009) and can thus be used for compound annotation in GC-MS measurements.

3.4.9 MetaboLights

In the field of proteomics, researchers who publish their work are also demanded to publish their experimental data in public databases like PRIDE (Martens *et al.*, 2005; Vizcaíno *et al.*, 2009). For metabolomics, the MetaboLights online repository was established for the same reason (Steinbeck *et al.*, 2012; Haug *et al.*, 2013; Salek *et al.*, 2013b). At EMBL-EBI it tries to collect metabolomics experiments raw data together with structured meta data.

[1] http://www.massbank.jp/en/statistics.html - accessed 04.02.2018
[2] http://mona.fiehnlab.ucdavis.edu/ - accessed 04.02.2018

3.4.10 Metabolomics Workbench

Established by the Data Repository and Coordinating Center (DRCC) of the National Institute of Health (NIH), the Metabolomics Workbench mainly takes the same line as MetaboLights (see above), but also provides a set of tools for statistical analyses (Sud *et al.*, 2016).

3.4.11 PredRet

The PredRet repository published by Stanstrup *et al.* (2015) tries to unlock the advantages of community-shared retention times for LC-MS. By matching retention times of identified compounds between different chromatographic systems, for new chromatographic systems predictive models based on the log P of other compounds can be established (log P is the logarithm of the partition coefficient, here a measure for lipophilicity.)

3.5 Summary

The here presented list of tools for metabolome informatics is large but of course still incomplete. The vast multitude of available (open source) software reflects the diversity of experimental, technical, and chemical approaches to the metabolome. Yet, the landscape of tools did not cover the necessity for comprehensive LC-ESI-MS spectra deconvolution - especially not with respect to SIL-aided isotopomer ratio analysis. ALLocator provides means for these aspects within a web platform, covering raw data preprocessing, highly interactive data exploration and curation (IDEC), and a thorough metabolite annotation workflow. ALLocator is presented in chapter 5. MeltDB (Neuweger *et al.*, 2008) was the first of the large GC-MS web platforms and serves as a one-stop-shop, serving from raw data upload until statistical data exploration. MeltDB 2.0, presented in chapter 7, introduces extensions in terms of data mining (unsupervised and supervised machine-learning) and IDEC.

4 Challenges for computational metabolomics

As soon as *wet lab* experiments are completed and measurements have been performed, the outcome relies on the joint success of computational tools and the operating analyst. Computational support is required (or at least extremely helpful) for every step from raw data acquisition, to data preprocessing, to data integration, to data analysis, to data exploration. The ultimate goal of the entire workflow (including *wet lab* procedures) is to gain a dataset and to derive new knowledge from it. The data flow and the single steps that need to be addressed by computational metabolomics are depicted in Figure 4.1 and will be described in more detail in the following sections.

Raw data processing is not the first time informatics come into play in metabolome analysis. Complex *wet lab* experiments ideally begin with computer-aided experiment design. The software-controlled parameterization of the data acquisition in the chromatograph and mass spectrometer have a very high impact on the resulting data and their analysis as well. However, in the context of this thesis, the computational data analysis workflow of interest begins after the raw data acquisition.

The first two categories of tasks shown in Fig. 4.1, preprocessing and integration, may not be underestimated but have been explained in detail in sections 2.4.2 and 2.5. Foremost the deconvolution of ions though is still a difficult matter for solely algorithmic approaches, as it is rather the rule than an exception that deconvolution results remain ambiguous. Not only are misinterpreted ion peaks a potential source of false positive annotations, but additionally may misinterpretations of the relations between ion peaks obfuscate the true neutral masses of metabolites and thus lead to false negatives at the same time.

Additional complexity comes into play through SIL experiments (cmp. section 2.6.5). The mixture of normal with U-^{13}C labeled samples in one analysis theoretically has the potential to double the number of monoisotopic peaks (U-^{12}C and U-^{13}C combined). Also, many of the isotopic patterns and respective mirrored isotopic patterns of ions with few carbon atoms will overlap, which hampers deisotoping.

Once deisotoping and ion deconvolution have been performed and pseudo spec-

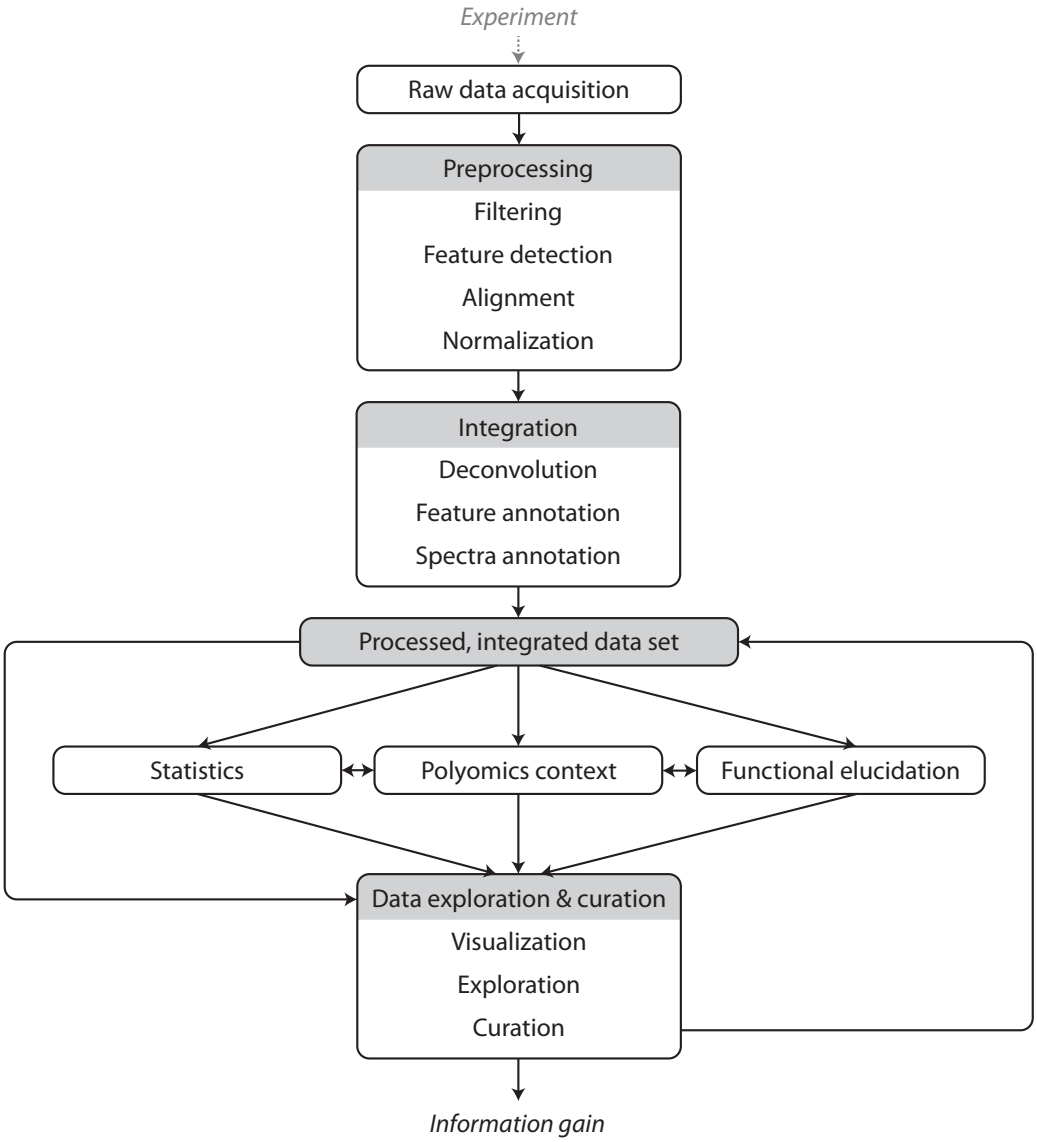

Figure 4.1: Tasks for computational metabolomics from raw data acquisition to curated results data sets. The steps are divided into four major categories: data preprocessing, data integration, data analysis, and interactive data exploration and curation. In practice, workflows do not necessarily have to cover every single step, but all major categories.

tra have been allocated from all the peaks, these spectra can be subjected to annotation strategies. While the very reproducible spectra (and also retention times, or at least retention indices) derived from GC-EI-MS may be matched against spectral reference databases (cmp. 2.5.2), LC-ESI-MS spectra rather have to be interpreted. LC retention times vary so wildly between the myriads of possible chromatographic methods that no public databases exist and the concept of retention indices could not be established on a community-wide basis. When using the exact same methods though, as for example defined by standard operating procedures, retention times are reproducible and can help for dereplication. The annotation of LC-ESI-MS spectra is limited to mass decomposition (cmp. section 2.5.3) and database matching by exact neutral masses or molecular formulas (cmp. section 2.5.4). The higher the mass of a molecule, and the lower the mass accuracy, the larger is the list of candidate molecular formulas for a single spectrum according to mass decomposition. Consequently it becomes more important to filter molecular formulas by more rules than just mass and mass deviation.

When quantitation tables have been generated after peak detection, quantitation and integration, these can be subjected to statistical analysis. Typical work horses of statistical analysis are t-test, ANOVA, PCA, PLS(-DA) and clustering methods. With the increasing wealth of data, answers to other question came into reach and naturally other statistical problems came into focus: Now machine learning methods like e.g. SVM, RF, and knn could be trained to classify new, yet unknown data sets. All these tools are described in literature and implementations are freely available for virtually all programming languages. But there are remaining challenges for statistical analysis that are more specific to the nature of metabolomics experimental data:

Still challenging are for example batch-effects, which hamper the integration of larger data sets through strong retention time shifts, changing matrix-effects, varying contaminants and others. These require tools and strategies for normalization and missing value imputation. Otherwise statistics, and consequently the significance of a result, can not be trusted. *Vice versa*, subtle changes in metabolite levels for different levels of a factor would not be detected, if these levels would not match the different batches.

Furthermore, there is a paradigm shift away from single biomarkers towards panels of biomarkers. Taking disease markers as an example, such biomarker panels can not only reveal a deranged metabolism but may also track back the cause of the distortion in a metabolic pathway and may thus become basis for the selection of a proper therapy (German *et al.*, 2005).

Such a combination of discovery science and functional elucidation through hypothesis-driven research within a metabolic network, but also brought into context with the genome, transcriptome, and proteome, is one of the mandates of sys-

tems biology (Ideker *et al.*, 2001).

Visualizations for the exploration of most statistical results or even plain metabolite levels is not a metabolomics-specific task. But both are abstractions and have come a long way from the initial raw data. Metabolomics software users ask for interactive visualizations to trace the entire process, including aspects like retention time alignments or ion deconvolution. This is not only important for the validation and quality assessment of results, but also for the possibility to revoke or curate parts of it. Neither the term 'exploratory data analysis (EDA)' nor 'information visualization (InfoViz)' nor 'interactive and dynamic visualizations' embodies the scientists immediate, active manipulation of data. These terms fail to express that after data exploration and quality assessment there should be options for quality improvement, be it manual or semi-automatic. The combination of modern information visualization and (manual) data curation might be best comprised with 'interactive data exploration and curation (IDEC)'.

5 ALLocator 1.0

LC-MS, especially in combination with ESI, is a major corner stone of metabolomics research for several years now and its applications today outnumber those of GC-MS. But it is not the data acquisition that is posing the greatest challenge to metabolomics: In a survey from 2009, asking for the greatest bottleneck of metabolomics, 35 % of the respondents named the identification of metabolites the biggest challenge, 22 % thought that assigning biological significances is most important, and 14 % decided that data processing/reduction is the crucial bottleneck (Milgram and Nordström, 2009). Six years later, an almost unaltered review of metabolomics bottlenecks was reported by Sévin *et al.* (2015).

The identification of truly novel compounds is not possible by MS alone, but requires complementary analytical techniques such as NMR (Scheubert *et al.*, 2013). Metabolite identification in the context of MS based metabolomics rather means assigning possible known molecular entities to all detected peaks, or better to peaks of interest. Using electrospray ionization, peaks can be observed representing so called pseudo-molecular ions. Here, intact analytes build adducts with small ionic species. Determining m/z-values with high accuracy allows the determination of a reasonable number of possible molecular formulas for each adduct by mass decomposition 2.5.3. Previous recognition of the type of adduct ($[M+H]^+$, $[M+Na]^+$, etc.) supports narrowing down the list of candidates.

In order to find biological meaning in the data, peaks must be associated to metabolites. This requires data reduction, i.e. filtering of noise, but also elucidation of peak interrelations: Molecules form adducts and fragments during LC-ESI-MS analyses. All these, as well as their isotopologues, generate distinct mass signals, but they are still representatives of the same original metabolite. The allocation of peaks that derive from the same analyte is called spectra deconvolution and was introduced in section 2.5.1. The artificial, purified spectra that contain all peaks that originate from the same single metabolite are called pseudo spectra.

Such a pseudo spectrum might for example comprise the peaks of the hydrogen ion adduct and the ion fragments created through the losses of water or ammonia ($[M+H]^+$, $[M+H-H_2O]^+$ and $[M+H-NH_3]^+$ respectively), which all derive from the same molecule M. Formed ions may even be ambiguous: For example, a mass difference of 17.027 Da can be explained by the neutral loss of ammonia ($[M+H-NH_3]^+$ and $[M+H]^+$) or by the formation of an ammonium adduct ($[M+H]^+$and $[M+NH_4]^+$). This is highly dependent on many technical parameters, for example

mobile phase composition and ion optic settings (see 2.3.6 Electrospray ionization (ESI) ion source for LC-ESI-MS).

One additional major aspect of LC-MS based metabolomics that has only recently stepped into the focus of cheminformatics is metabolite quantitation via isotope labeling. The use of stable isotope labeling (SIL) has become an important and popular approach in the field of metabolomics. Many strategies using SIL were developed, enabling more accurate metabolite identification and quantitation in complex biological samples (Mashego *et al.*, 2004; Baran *et al.*, 2010; Bueschl *et al.*, 2012). The numerous advantages of this common approach have been reviewed in Bueschl *et al.* (2013a). Common to most SIL experiments is the mixing of naturally labeled (unlabeled) samples with samples that are enriched with stable isotopes and the analysis of these mixed samples by GC- or LC-MS. Either one group of samples from one experimental condition is unlabeled and another set of samples from a second experimental condition is labeled, or both groups of samples are unlabeled and a labeled internal standard is added to each sample. In any case, this allows calculating abundance ratios of metabolites in the two samples, while matrix effects can be neglected (Mashego *et al.*, 2004; Bueschl *et al.*, 2012). Additionally, the distance between the signals of the unlabeled and fully labeled isotopologue peaks provides substantial benefits for metabolite identification as it can be used to infer the correct number of atoms of the respective element in the analyte. This adds a powerful filter for false positives to molecular formula generation.

The broad range of available separation methods (i.e. columns, solvents and gradients) allows for data acquisition of small molecules with a great coverage. The data output is complex and it takes a number of steps to process, integrate and interpret the raw signals. A comprehensive overview of these steps is given in Fig. 4.1. The current landscape of metabolomics software provides solutions for each step of the entire processing from LC-MS raw data, to signal processing, to metabolite identification and relative quantitation. Nevertheless, it misses one that (a) uses the full potential of ^{13}C-stable isotope labeling for metabolite and fragment annotation, (b) is optimized for mass isotopomer ratio analysis, (c) provides users with an interactive interface not only to explore but also to modify the results of automatic processing (see chapter 4), and finally (d) addresses the strong and well advanced evolution of research projects towards cross-group collaborations (Wuchty *et al.*, 2007). To fill these gaps we developed the ALLocator system, presented in this manuscript. ALLocator is a novel web-platform particularly for the comprehensive analysis of metabolomics LC-ESI-MS (labeling) experiments and is streamlined for mass isotopomer ratio analysis. It covers all aspects (a) - (d), as shown in the application example in chapter 6.

The following sections are dedicated to present the ALLocator web platform in detail. Especially the new ALLocatorSD algorithm for spectra deconvolution

will be explained in depth. An additional section will furthermore clarify how the web platform allows for interactive data exploration and curation (IDEC). Fig. 5.1 comprises the ALLocator workflow from peak detection to IDEC and finally data export in a concise scheme.

5.1 System design and implementation

The described scope of the ALLocator web platform, focused on LC-ESI-MS-centered metabolomics, led to a number of high-level requirements which guided the development of this software:

- Support researchers beginning from raw data upload.
- Allow to share data and to collaborate on data.
- Provide freely parameterizable preprocessing pipelines that cover all steps that can be reliably automated.
- Give deep insight into the preprocessing results.
- Empower the user to freely curate, refine and enhance the preprocessing results.
- Integrate the results with knowledge from public repositories.

The ALLocator web platform comprises methods and tools for the semi-automated analysis of LC-ESI-MS experiments, from the import of chromatographic raw data, to the export of lists of annotated and quantified compounds. Users can create experiments and upload chromatograms (CDF-, mzXML-, mzData files) of both positive- and negative-mode measurements. The web interface then guides the user through the customizable pre-processing steps, and finally displays the results in interactive and dynamic visualizations for data exploration and manual annotation. The general concept of all features is to achieve transparency of the data, i.e. to provide researchers with all information to support decisions in peak annotations, rather than to plot irrevocable results of black box algorithms.

5.1.1 System integration

The ALLocator system integrates a number of platforms and tools to provide its features to the user. The following paragraphs are dedicated to explain the technical aspects and software design decisions that have driven the development of the ALLocator web application.

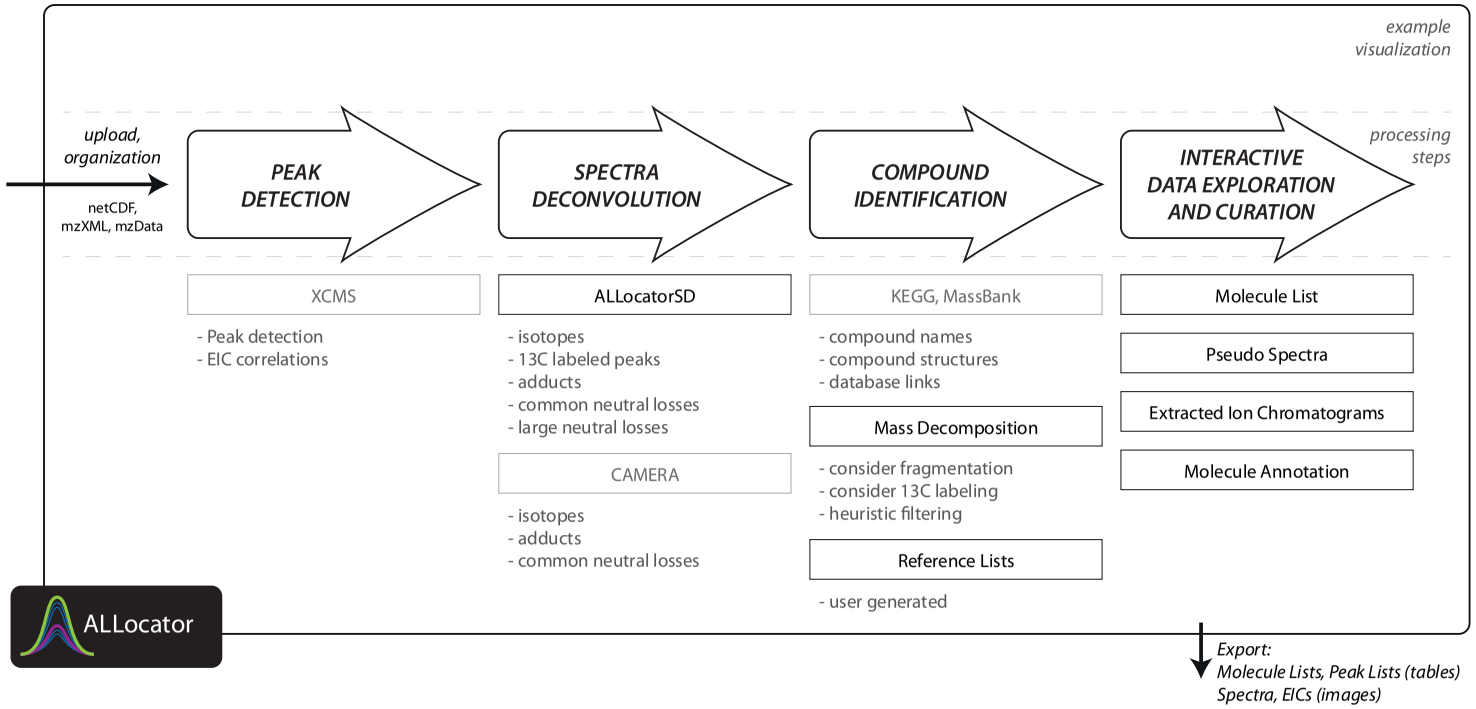

Figure 5.1: The overview shows the processing steps in ALLocator, including peak detection, spectra deconvolution, compound identification, and interactive data exploration and curation. For each step the provided tools are listed. Third-party tools are displayed in a grey color. Data can be exported as peak or molecule lists. Spectra and extracted ion chromatograms can be exported in various image file formats.

Platform

The main parts of the software were developed using Java, applying Spring[1], Hibernate[2], and MySQL[3] for the platform setup. R-based third party tools were integrated using the RCaller library (Satman, 2010). The web interface is delivered via Java Server Faces[4] and designed with help of the bootstrap CSS library[5]. Beyond that, interactivity and interconnectivity has been enhanced with JavaScript, and often jQuery[6]. The display of spectra and chromatograms was realized with the Highcharts JS library[7].

As ALLocator is part of the CeBiTec bioinformatics software platform, projects, users and restricted access are managed by the General Project Management System (GPMS).

Preprocessing tools

XCMS and CAMERA are both third-party R tools. XCMS was integrated to perform the initial peak picking, the mandatory first step in any ALLocator preprocessing workflow. The ouput of XCMS (peak lists) is used as an input for the second step: spectra deconvolution. To offer an alternative to the ALLocator own spectra deconvolution tools (see section 5.3), CAMERA was integrated as well. In fact ALLocator was designed to explicitly support CAMERA results in its own data model. Vice versa, ALLocatorSD results can be exported in a format that fits the CAMERA text output. This facilitates exchangeability of the two spectra deconvolution tools in existing preprocessing workflows.

Metabolite identification tools

ALLocator integrates metabolite databases in different ways. The KEGG COMPOUND database (Kanehisa and Goto (2000), version of 2010) is included in the ALLocator database for efficient search by monoisotopic masses. The internal KEGG searches are automatically triggered after spectra deconvolution and associates metabolites to pseudo spectra. The MassBank (Horai *et al.*, 2010) spectral database is accessed via its SOAP web service. Pseudo molecular ions and their fragments can be matched against MassBank MS/MS spectra in order to easily create manual annotations. And last, molecular formulas that result from mass

[1] https://spring.io/
[2] http://hibernate.org/
[3] https://www.mysql.com/
[4] https://javaee.github.io/javaserverfaces-spec/
[5] https://getbootstrap.com/
[6] https://jquery.com/
[7] https://www.highcharts.com/

decompositions are linked to the respective search pages in ChemSpider (Pence and Williams, 2010).

5.1.2 Data model

The design of the data model was mainly driven by the following three key rationales:

1. The system has an extensible set of tools that create results from input data
2. The parameters for each tool can be set by the user
3. Results of spectra deconvolution tools retain all information and remain modifiable

See in Fig. 5.2 how these key aspects were addressed in the relational database model. The user is capable of creating experiments, uploading chromatograms, and running tools with the parameterization of choice. That fulfills the first two aspects. However, the most important rationale is the third one: By designing an annotated many-to-many relationship between pseudo spectra and peaks, the often ambiguous results of spectra deconvolution can be persisted and remain accessible for curation. This allows a system, in which the user is always in control of the result and can decide which pseudo spectra are correct and which are misinterpretations. Refer to section 5.5 to find an illumination of this topic from a graphical user interface perspective.

The model for the identification of metabolites was designed following the same idea: Many metabolites can be associated to a pseudo spectrum (if they have identical or very similar monoisotopic masses), but the user is in charge of confirming the correct one. Section 5.4.2 provides more information on automated database search by masses. To not only rely on the monoisotopic masses, but to make use of the specific fragmentation patterns of metabolites, the user may also create and persist personal reference lists of pseudo spectra (cmp. 5.4.5).

5.1.3 User management

As stated earlier, ALLocator is connected to the General Project Management System (GPMS) of the CeBiTec Bioinformatics Platform. All ALLocator users have to log in with their CeBiTec GPMS accounts. Beyond that, certain objects in ALLocator are subject to access control lists (ACLs). These allow the owners (i.e. creators) of experiments and reference lists to share their results with fellow scientists by granting access to their respective GPMS accounts. These access permissions can eventually be revoked by the owner.

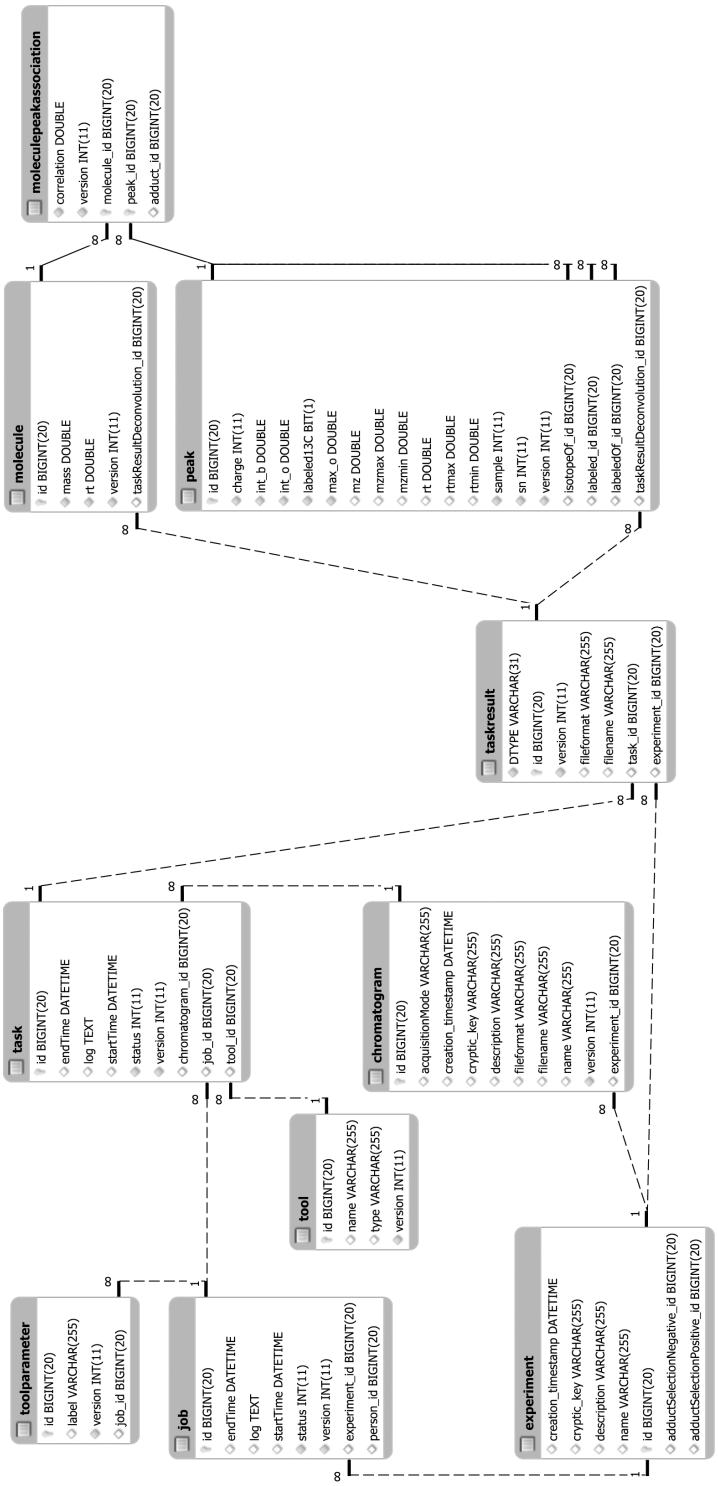

Figure 5.2: Excerpt of the ALLocator database model. The included tables are relevant for tools (XCMS, ALLocatorSD, CAMERA) that are run on an experiment (job) and all chromatograms in it (tasks). Taskresults refer to R result files in the file system and/or to lists of peaks and molecules.

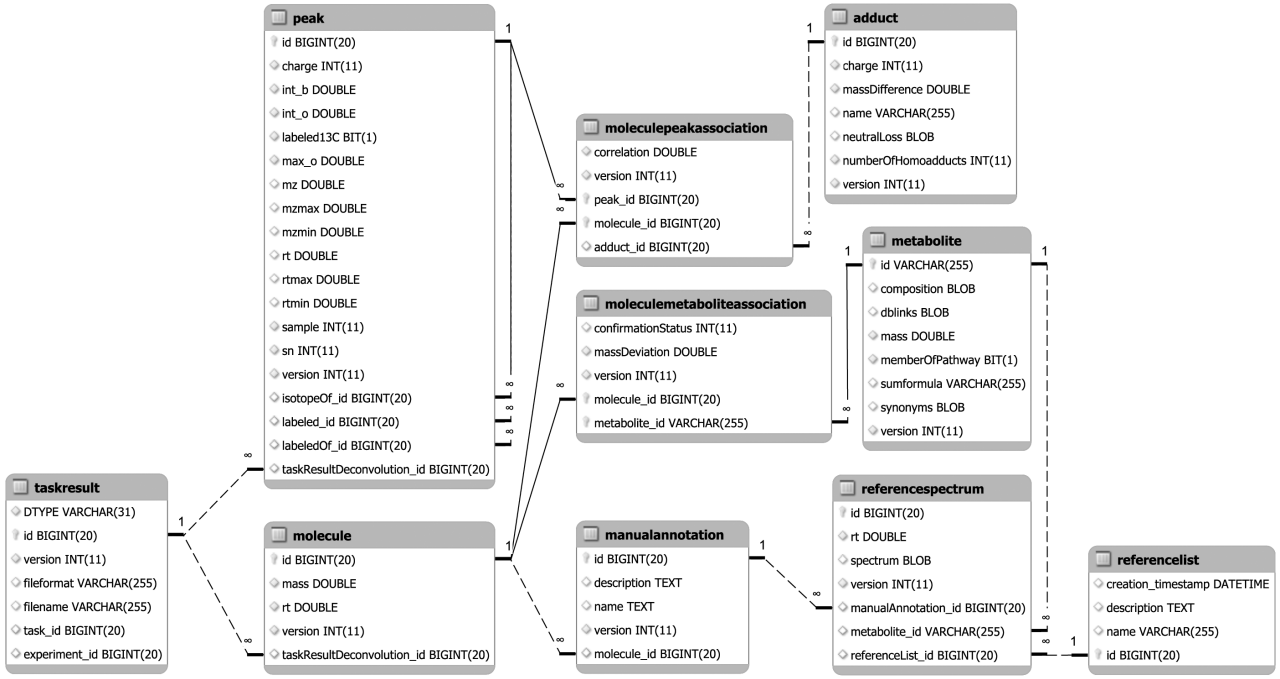

Figure 5.3: Excerpt of the ALLocator database model. The included tables are relevant either for the relationship of peaks and molecules, or for the annotations of molecules.

5.2 Pre-processing methods

Pre-processing algorithms that are offered by ALLocator can be started either for a single chromatogram or for all the chromatograms of an experiment at once. Users can set parameters for these algorithms through the web interface. These pre-processing jobs are submitted to the compute cluster of the Center for Biotechnology of Bielefeld University (CeBiTec), hosted by the Bioinformatics Resource Facility (BRF). Whenever the Java software has to call programs running in the R environment (R_Development_Core_Team, 2011) (version 2.13.2), this is realized through the Runiversal package (Satman, 2010) for R.

In the ALLocator workflow (see Fig. 5.1), the first job to execute applies the centWave (Tautenhahn *et al.*, 2008) LC-MS feature detection method of the XCMS (Smith *et al.*, 2006) software (version 1.26.1) for R. Generated peak tables and the R object are stored to serve as input for the next step in the workflow: spectra deconvolution. Now, two options are available: either the new ALLocatorSD algorithm for spectra deconvolution, which will be described in detail in the next section, or the CAMERA tool for compound extraction and annotation (Kuhl *et al.*, 2012; Kuhl and Tautenhahn, 2010). Both use the XCMS results as input, and both generate output that can be explored and processed manually using the visualizations and user interface of the ALLocator web platform (see section 5.5).

5.3 Spectra deconvolution algorithm

One of the core features of the software platform is 'ALLocatorSD', the new algorithm for mass spectral deconvolution in LC-ESI-MS data. Its paramount objective is to facilitate the interpretation of convoluted spectra by grouping peaks that can be associated to the same potential molecule. To this end, peaks have to be set into relations between each other. This exacts to identifying peaks from isotopes, adducts and fragments which may have derived from the same original metabolite with a monoisotopic mass of M. This procedure is explained below by splitting it into seven steps. Here, all steps are described as applicable for a chromatogram measured in positive mode. However, this approach can be applied for negative mode measurements without any restriction. Step four and six are only applicable to U-^{13}C SIL experiments.

For the ease of reading, please note two requirements that are made for any two peaks to be compared:

- In all steps, peaks are only compared to each other if the deviation of their retention times (RT) is less than ε_{RT}, the allowed retention time error as defined by the user.

- Monoisotopic masses and m/z values are considered to be equal, if they are within the user-defined accepted mass-to-charge error $\varepsilon_{m/z}$.

Step 1 - read peak lists: The ALLocatorSD pipeline annotates and groups peaks that were identified by XCMS. Thus, the first step is to read in the peak lists generated by this R tool. Each peak is, among other attributes, defined by a RT and m/z value.

Step 2 - identify isotopes: Many peaks in each measurement do not represent monoisotopic molecules, but one of their isotopologues. It is important to identify and annotate them as such for a couple of reasons:

First, annotating them as isotopologues ensures that they are not misinterpreted as monoisotopic peaks in a later step. This can be seen as a step of data reduction.

Second, isotopologues reveal the charge of a molecule. The m/z of isotopologues from a single-charged molecule will increase by 1.003355 Da, each. Is the molecule charged twice (i.e. $z = 2$), the m/z distance between the isotopologues will be half of that.

Third, in case of a ^{13}C-labeling experiment this very same step but with decreasing m/z values is applied to identify monoisotopic ^{13}C peaks by their lighter isotopologues. Throughout this thesis, these will be called *mirrored isotopes*, as the isotope patterns of incompletely ^{13}C-labeled molecules resemble a mirrored version of the molecules natural isotopic patterns.

Step 3 - identify common adducts and neutral losses: Users can choose from a predefined list which common adducts and neutral losses they expect or assess possible in their measurements. From all selected ions a distance matrix can be created. Any two monoisotopic peaks eluting at the same RT with a mass difference found in that matrix can then be annotated as the corresponding ions. Not all adducts and neutral losses are equally frequent though. While, for instance, any valid spectrum of a metabolite may be expected to feature peaks for at least $[M+H]^+$ or $[M+Na]^+$, other ions like $[M+H\text{-}C_6H_{12}O_6]^+$ are less common and for most metabolites not even possible. To account for this, users may define any number of ions as 'seeds': Only pseudo spectra that contain at least one seed will be allocated. As a positive side effect, this restricts the elements of the distance matrix that have to be considered from the entire strict upper matrix to a smaller subset. This effect can be seen in the equations (5.1) and (5.2). The default list of common adducts and losses for ALLocatorSD features 23 single-charged ions, of which three are defined as seeds. This results in 63 unique distances (as opposed to 253).

$$\frac{i(i-1)}{2} \tag{5.1}$$

, will be restricted to

$$\frac{s(s-1)}{2} + (i-s)s \tag{5.2}$$

i number of all selected ions
s number of ions that where defined as seeds

Any set of adducts and fragments, including their isotopologues, that putatively derive from the same original molecule with a mass M, is allocated to form a 'pseudo spectrum' for M at the given RT.

Step 4 (SIL only) - associate U-^{13}C peaks: In step 2 a number of peaks may have been identified to be U-^{13}C peaks, because of their *mirrored isotopic patterns*. In step 4 these are associated to their ^{12}C counterparts. To be annotated as the ^{13}C monoisotopic peak of another ^{12}C monoisotopic peak, the former has to feature a mass that is $n \times 1.003355$ higher than the latter, where n is a positive, natural number. As n is the number of carbon atoms in both ions, it must be in the range of possible occurrences of carbon atoms in the ion according to mass decomposition. Later, in return, only mass decompositon results with an exact number n of carbon atoms will be accepted as valid results. As an example, the associated ^{13}C monoisotopic peak of a pseudo molecular ion $[M+H]^+$will consequently be annotated as $[M+n+H]^+$.

Step 5 - identify homoadducts: The fifth step aims to identify *homoadducts* (sometimes also referred to as *multimers* or *multi-masses*). *Homoadducts* are two moieties of the same analyte that are attached to each other before ionization. The most frequent example is $[2M+H]^+$. Finding these is obviously easy, as soon as M is known. This prerequisite is fulfilled for all pseudo spectra as of step 3.

Step 6 (SIL only) - identify large, uncommon neutral losses: The information on the number of carbon atoms n in ^{12}C/^{13}C peak pairs that was assessed in step 4 can be used to identify fragments with large neutral, uncommon losses that were not predefined. Step 6 may thus complement step 2. If for the primary adduct p of a pseudo spectrum a pair of ^{12}C/^{13}C peaks exists (e.g. the $[M+H]^+$ and $[M+n_p+H]^+$), a requirement for all molecular formulas s_p in the list S_p of respective mass decompositions is to feature n_p carbon atoms. If a potential ^{12}C/^{13}C fragment peak pair f with a comparatively lower m/z-values exists at the same retention time and if it was not yet assigned to another pseudo spectrum, a list of possible molecular formulas S_f with each having n_f carbon atoms is calculated. The mass difference between the peak pairs p and f is substituted to mass decomposition in order to elucidate the neutral loss l, and returns a list of possible molecular formulas S_l with a required number of carbon atoms $n_l = n_p - n_f$. If in S_p, S_f

and S_l a unique triplet of molecular formulas s exists which satisfies $s_l = s_p - s_f$, the fragment peak pair f can be annotated as the neutral loss [M+H-s_l]$^+$ of the same M as the [M+H]$^+$.

Step 7 - validate peak correlations: The last step is to validate all pseudo spectra by checking whether the peak's intensities correlate well over time. Any peak that does not correlate well enough to the primary peak - i.e. the correlation is worse than a user-defined threshold - will be discarded from the pseudo spectrum.

After completion of all steps the result is a list of annotated peaks and pseudo spectra. Each pseudo spectrum contains adduct- and fragment-peaks that represent a common monoisotopic mass M. Isotopologues are associated to their monoisotopic peaks, and monoisotopic ^{13}C peaks are associated to their monoisotopic ^{12}C counterparts. Monoisotopic peaks that could not be allocated by any pseudo spectrum may nevertheless be associated to their isotopologues and can still be identified as U-^{13}C peaks.

5.4 Metabolite annotation

The different levels of metabolite identification were introduced in the background section 2.5.4. Now the term 'metabolite annotation' describes the assignment of a name and/or molecular formula to one or more measured signals - or in this context more specifically - to a set of peaks in an LC-ESI-MS measurement. Metabolite annotations can refer to any identification level and often carry additional information such as database identifiers. The section at hand is dedicated to the means for metabolite annotation which are implemented in the ALLocator software. Fig. 5.4 shows a table of pseudo spectra that were found in a measurement. Note that only metabolite names that are decorated with a green check mark ('confirmed') are considered as metabolite annotations. Confirmed metabolite annotations can be created via completely manual annotations, via prepopulated manual annotations, by confirming KEGG Compounds, or by applying custom reference lists. These are explained in the following subsections, using the spectrum of L-citrulline as an example.

5.4.1 Manual annotation

Manual annotations are a simple way to add a generic identifier to a pseudo spectrum. The user can define a name and a description to any pseudo spectrum. These very basic annotations allow to reidentify the same compounds in multiple analyses (cmp. subsection 5.4.5) and make them accessible for statistics.

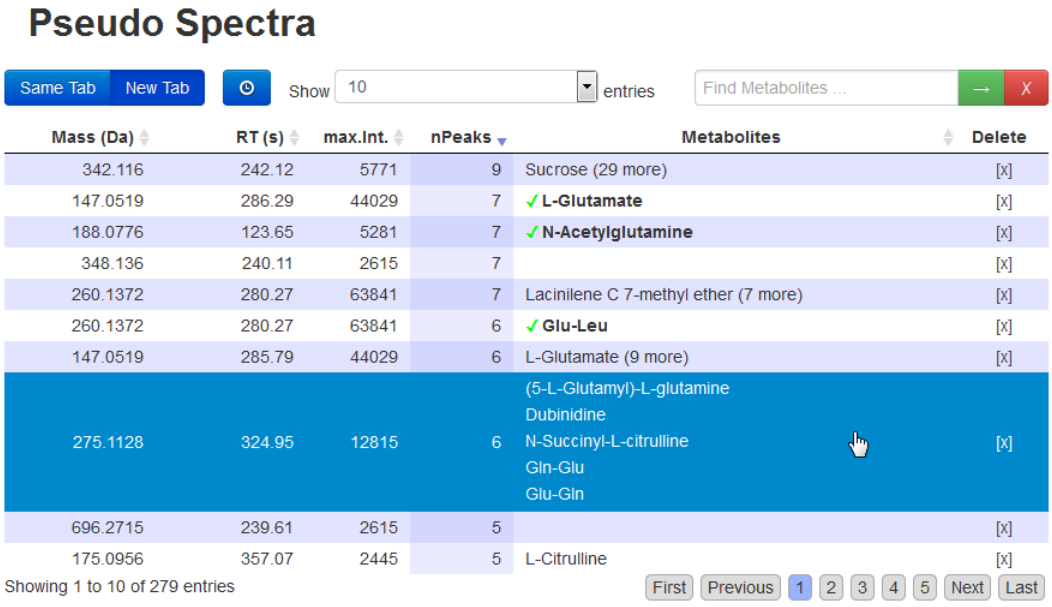

Figure 5.4: A part of the Pseudo Spectra page in ALLocator. The screenshot shows a subset (ten entries) of all 279 pseudo spectra that were found in the given analysis. The table presents the monoisotopic mass (Da), the retention time (s), the intensity of the highest monoisotopic peak, the number of associated monoisotopic peaks, and the associated metabolites. The latter is either a number of KEGG Compounds with similar monoisotopic masses (e.g. Sucrose and 29 more candidates) or a confirmed annotation (e.g. L-glutamate). Hovering a table cell like '(5-L-glutamyl)-L-glutamine (4 more)' will reveal the complete list of KEGG Compound candidates. Annotations can be confirmed by manual annotations, by using custom reference lists, or by confirming a KEGG Compound. Confirmed annotations are indicated with a green check mark and a bold font face.

5.4.2 Search by monoisotopic mass (KEGG)

A KEGG Compound search is integrated into the ALLocator preprocessing pipeline, whether the user chooses ALLocatorSD or CAMERA for spectra deconvolution. Pseudo spectra are linked to all KEGG Compounds with the same monoisotopic mass (within the user-defined mass tolerance ε_m) which are called 'metabolite candidates' in the following. The pseudo spectra page allows to search pseudo spectra that fit a certain metabolite candidate (cmp. 'Find Metabolite ...' in Fig. 5.4). The 'Metabolites' tab of a pseudo spectrum page reveals more details on each metabolite candidate. Here, the user can also confirm a metabolite candidate, which creates a confirmed metabolite annotation (see Fig. 5.5).

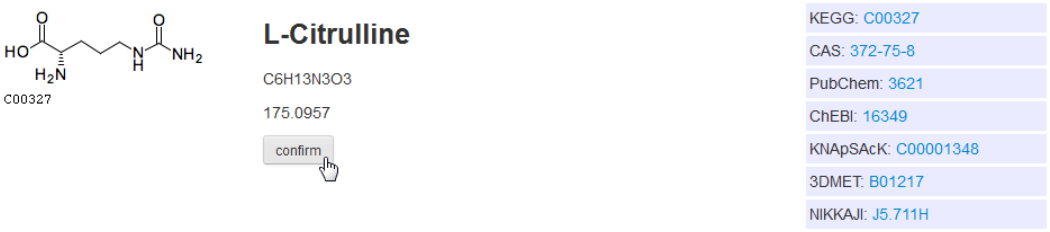

Figure 5.5: The Compound tab in ALLocator. Searching KEGG for the monoisotopic masses $\vec{M}$ of all pseudo spectra that were generated during spectra deconvolution is part of the ALLocator preprocessing pipeline. When reviewing a pseudo spectrum, the user can open the Compound tab to see a list of KEGG (Kanehisa and Goto, 2000) Compounds that match the measured monoisotopic mass M of the putative molecule (within user-defined allowed mass deviation ε_m). The Compound tab presents all hits along with their name, molecular formula, monoisotopic mass, their molecular structure and external hyperlinks for connected database identifiers. Confirming a KEGG hit makes it the metabolite annotation for the pseudo spectrum.

5.4.3 Search by spectrum (MassBank)

The MassBank (Horai *et al.*, 2010) spectral library contains mostly MS/MS spectra and is thus not feasible for all pseudo spectra in ALLocator. As ALLocator is mainly tailored to LC-ESI-MS measurements, not all pseudo spectra feature sufficient fragment peaks. However, if in-source fragmentation has occured and at least a few neutral losses can be observed, ALLocator allows to mimic MS/MS spectra that only include the pseudo molecular ion ($[M+H]^+$ in positive mode or

Figure 5.6: The MassBank tab in ALLocator. If pseudo spectra contain sufficient fragment ions ALLocator allows to use them for simulated MS/MS spectra and to match them against the MassBank MS/MS spectral library (Horai *et al.*, 2010). The presented results include the spectrum identifier, MassBank identifier, the molecular formula, monoisotopic mass and match score. The information can be used to create a manual annotation for the pseudo spectrum.

$[M-H]^-$ in negative mode) and all its fragments. Adducts and multiply charged ions are ignored. These mimicked MS/MS spectra can be matched against the MassBank MS/MS spectral library in order to find new metabolite candidates or to validate KEGG metabolite candidates. Results of the MassBank search are presented in a list format similar to the KEGG Compounds, as can be seen in Fig. 5.6. The user can easily create manual annotations that are prepopulated with information from the selected MassBank hit.

5.4.4 Mass decomposition and search by molecular formula (ChemSpider)

The mass decomposition tool that is integrated into ALLocator generates lists of elemental compositions (i.e molecular formulas) that fit the monoisotopic mass M_{PS} that was figured from a pseudo spectrum. For this, the user has to define an allowed mass error ε_M, such that every listed molecular formula will have a monoisotopic mass M_{MF} within the closed range $[M_{PS}-\varepsilon_M, M_{PS}+\varepsilon_M]$. The set of elements that is considered for mass decomposition, sometimes called the atomic alphabet, is limited to $L = \{C, H, N, O, P, S\}$, which is a reasonable assumption for small biomolecules, especially in a prokaryote context.

The theory of mass decomposition, as well as the growth of decomposition results with both M_{PS} and ε_M, is explained in the background section 2.5.3. The ALLocator implementation of mass decomposition however goes beyond that, by optionally reducing the number of generated candidates based on a set of rules that are built on two different foundations: biochemical knowledge and mass spectral information.

The first includes five of the rules that were suggested by Kind and Fiehn (2007). Some of these rules were defined after a statistical evaluation of public compound databases and the subsequent determination of boundaries for certain element occurrences. The authors also suggest to apply the LEWIS and SENIOR rules (Senior, 1951), which take the valence states of all atoms in a molecule into account. The last knowledge-based rule would only allow molecular formulas to contain phosphorous due to phosphorylations (Kessler *et al.*, 2014). These first six filters, based on biochemical knowledge, are explained in more detail below. Fig. 5.7 shows that activating these filters effectively reduces the number of formula candidates. Note again that these filters were evaluated using public databases and are designed to only exclude the most exotic molecular species, if at all (Kind and Fiehn, 2007; Kessler *et al.*, 2014).

Filter by element numbers This heuristic rule defines upper boundaries for the numbers of atoms of each element in generated formulas. These upper boundaries depend on M_{PS}. Boundaries were determined for molecules with masses below 500 Da, below 1000 Da, below 2000 Da, and below 3000 Da. E.g. a molecular formula generated for a mass below 1000 Da may not contain more than 78 carbon, 126 hydrogen, 25 nitrogen, 27 oxygen, 9 phosphorous, or 14 sulfur atoms (Kind and Fiehn, 2007).

Filter by element probability Also heuristic, this rule defines upper boundaries for the numbers of heteroatoms in a generated molecular formula, if multiple of them occur. E.g. if a formula contains more than three atoms of each

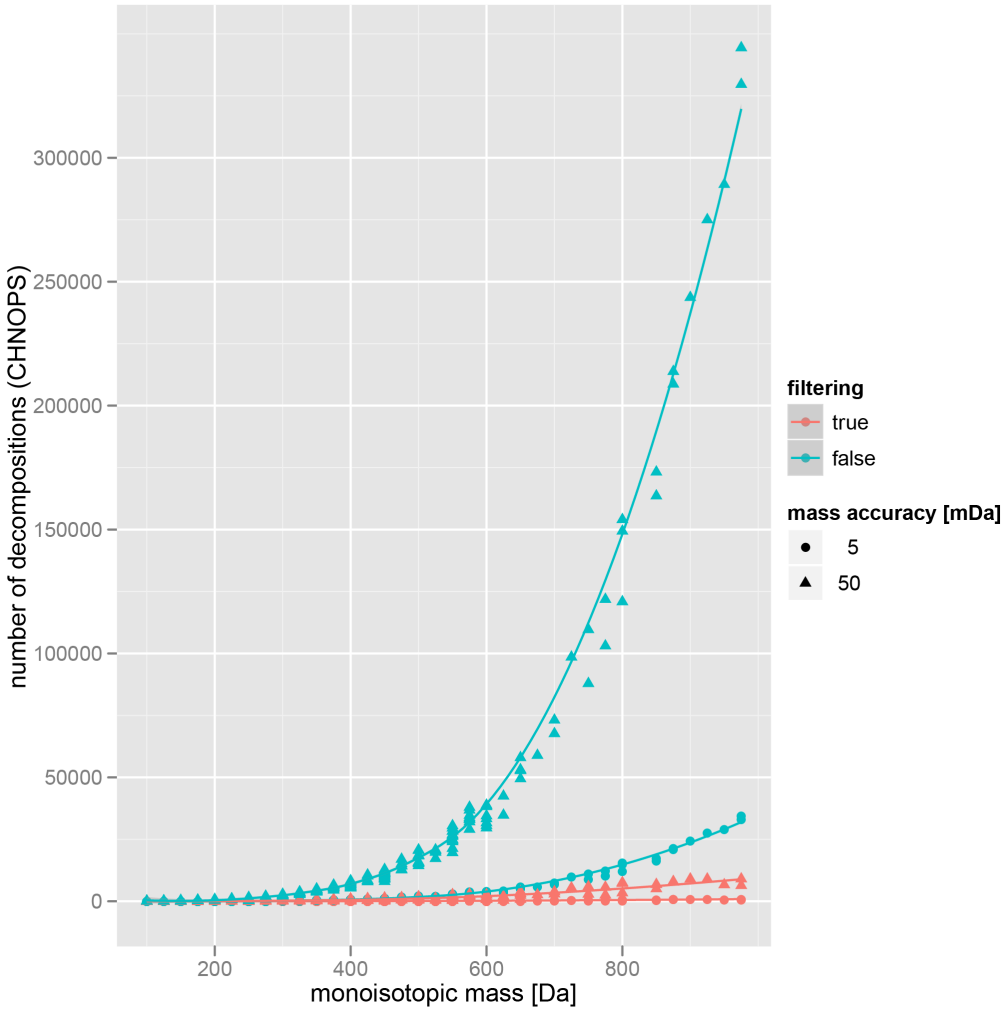

Figure 5.7: The number of calculated mass decompositions in dependency of the target monoisotopic mass, the mass accuracy, and applied filters for the atomic alphabet of $L = \{C, H, N, O, P, S\}$. For nominal masses from 100 Da to 1000 Da, in 25 Da steps, compounds from the KEGG COMPOUND database have been selected. For each of these compounds monoisotopic masses, decomposition with two different mass accuracies was performed (5 mDa and 50 mDa) using the mass decomposition and filtering algorithms of the ALLocator software.

element nitrogen, oxygen, and phosphorous, it may not contain more than 11 nitrogen, 22 oxygen, or 6 phosphorous atoms (Kind and Fiehn, 2007).

Filter by element ratio The third heuristic rule restricts the ratios of hydrogen to carbon and heteroatoms to carbon. E.g. 99.7 % of molecules below 1500 Da in the Wiley mass spectral database show an H/C ratio in the range of 0.2 to 3.1 and an O/C ratio below 1.2 (Kind and Fiehn, 2007).

Lewis rule Sometimes referred to as 'octet rule' it demands that the atoms of generated formulas can share electrons such that the *s*, and *p*-valence shells are completely filled. Then, for this formula a correct structure of a neutral molecule can be drawn and the species potentially exists. Validating this, the Lewis rule is fulfilled if sum of all valence electrons in the molecular formula is even (considering the highest possible valences for each element) (Kind and Fiehn, 2007).

Senior rule The Senior rule extends on the Lewis rule and furthermore requires that the sum of valence electrons is at least twice the valence of the element with the highest valence present in the formula. Furthermore, the sum of all valences must be at least twice the total number of atoms in the formula minus 1 (Kind and Fiehn, 2007; Senior, 1951).

O/P ratio $>$ 3.3 Motivated by the fact that phosphorous most commonly appears in biomolecules as part of an $O(PO_3)_n$ substructure, where n is between 1 and 3, this rule requires that per phosphorous atom at least 3.3 oxygen atoms are present (Kessler *et al.*, 2014).

In addition to these six biochemical knowledge-based rules, two more rules have been developed for ALLocator, which take the mass spectral information of the input pseudo spectrum into account:

Consider Pseudo Spectra Spectra deconvolution allocates all product ions into a pseudo spectrum that represents the original molecule with the monoisotopic mass M (see step 3 in subsection 5.3). Typically these can be a pseudo molecular ion, adducts, and/or fragments. Among them, fragments bear additional structural information which can be included into mass decomposition. To give an example, the fragment ion $[M+H-H_2O]^+$ indicates a neutral loss of water. It is thus clear that the molecular formula of the original molecule must contain at least one hydrogen and one oxygen atom. Similarly, the ion $[M+H-NH_3]^+$ proofs that the correct molecular formula for M must contain at least two hydrogen atoms and one nitrogen atom. The larger the neutral loss, the more revealing the ion is for the filter of molecular formula candidates (cmp. $[M+H-NH_3-CH_2O_2]^+$ or $[M+H-C_6H_{12}O_6]^+$): Every

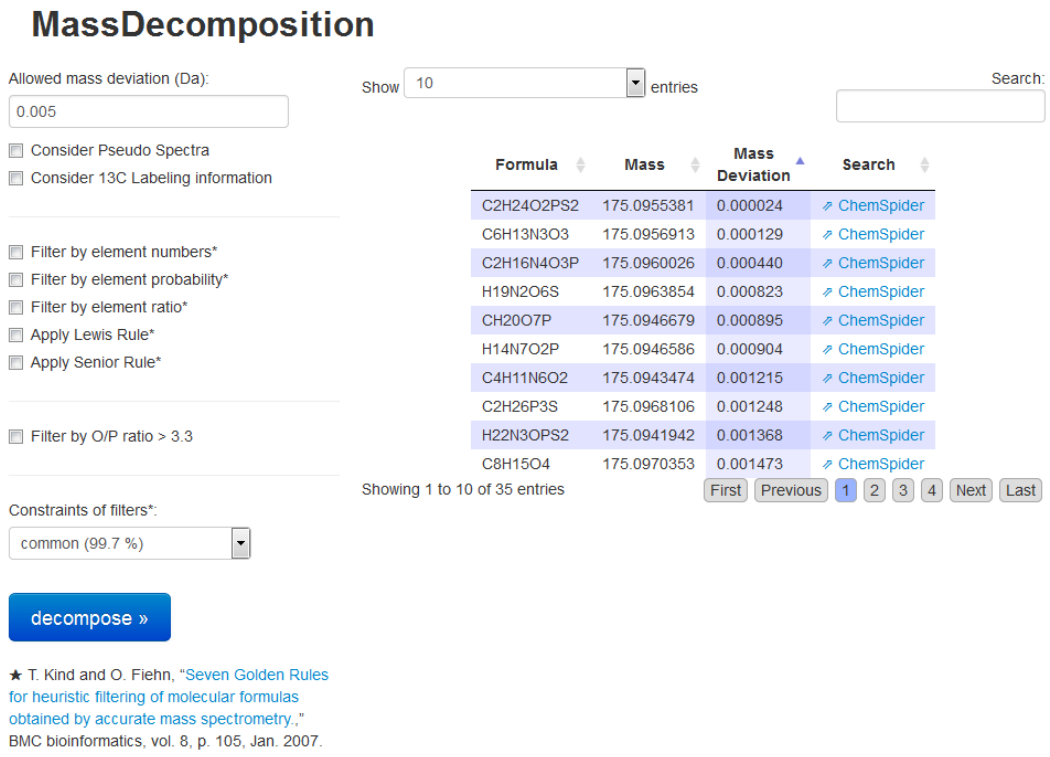

Figure 5.8: The MassDecomposition tab in ALLocator. The example shows unfiltered mass decomposition results for a $^{12}C/^{13}C$ spectrum of L-citrulline (measured monoisotopic mass $\sim$ 175.095562 Da) applying an allowed mass deviation of 0.005 Da. The unfiltered mass decomposition generated 35 molecular formulas. Activating only the 'Consider Pseudo Spectra' filter results in ten molecular formulas. Activating only all six chemical filters results in two molecular formulas. Activating only the 'Consider 13C Labeling information' results in the single molecular formula $C_6H_{13}N_3O_3$, which is the correct molecular formula of L-citrulline. Each presented molecular formula provides a hyperlink to a respective ChemSpider search, which allows to find compounds that fit the given formula.

generated molecular formula $s_{candidate}$ is discarded, if it does not contain each substructure s_n in the set of all observed neutral losses S_n. Note that these substructures are not summed up, as *a)* multiple subsequent fragmentation events can produce a series of ion species, and *b)* different observed neutral losses may 'overlap' in that they share certain atoms of the original structure.

This rules becomes extraordinary powerful in U-^{13}C SIL experiments, where large, uncommon neutral losses can be identified (see step 6 in subsection 5.3).

Consider ^{13}C labeling information This rule takes effect only, if in the pseudo spectrum a ^{13}C peak was identified for the pseudo molecular ion (see step 4 in subsection 5.3). Then, the number of carbon atoms n_C in the original molecule can be read from the integer distance of the monoisotopic masses of the ^{12}C and the ^{13}C peak. Every generated molecular formula $s_{candidate}$ is discarded, if it does not exactly feature n_C carbon atoms.

In fact, filters are not necessarily applied after mass decomposition. The latter two filters give a lower constraint to the search space before mass decomposition. The filters for maximum element numbers and element ratios can terminate the mass decomposition algorithm early and this way give an upper constraint to the search space.

Fig. 5.8 shows the ALLocator user interface for mass decomposition, where all filters can be opted for. Every generated molecular formula that remained after filtering is listed in a sortable result table and provides an external link to a ChemSpider search for the given formula. Both a selected molecular formula and additional information from ChemSpider (or other sources) can finally be taken to create a manual annotation for the pseudo spectrum.

5.4.5 Custom reference lists

Reference lists are a means for dereplication. After some effort has been invested to annotate pseudo spectra in one sample, the user can persist them as references. Each reference will consist of m/z/intensity pairs for all monoisotopic peaks, the putative mass of the original compound, the retention time of the pseudo spectrum, and the confirmed annotation for the spectrum. Applying this reference list to another sample will automatically allocate and annotate pseudo spectra that are similar to one of the references. That way, not only spectral annotations but also deconvolution results can be dereplicated. The similarity is determined using the cosine score as explained in subsection 2.5.2 and equation 2.9. Here, the intensity of each monoisotopic peak in the query pseudo spectrum is matched with

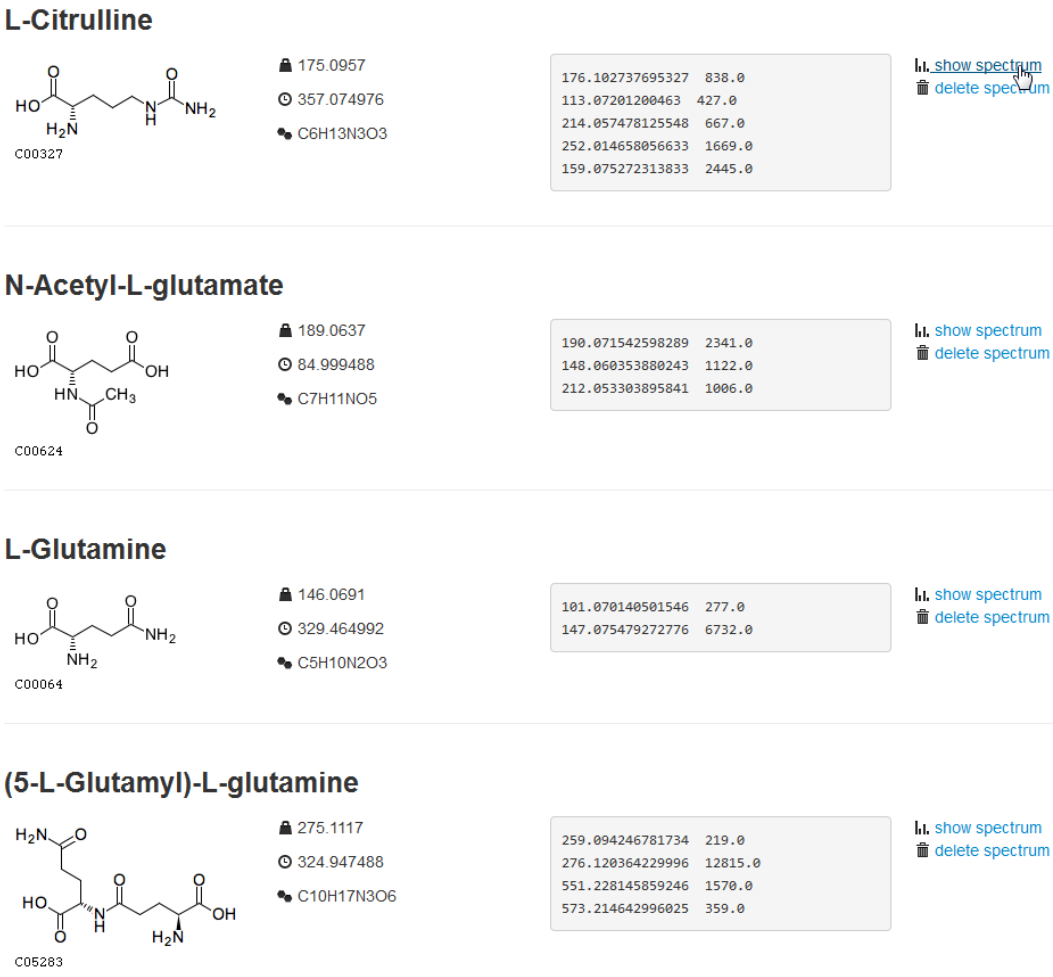

Figure 5.9: The reference list review page in ALLocator. The user can create reference lists from annotated measurements: All pseudo spectra with metabolite annotations will be taken as reference spectra and retention times for the given metabolite. This allows to easily apply the same annotations to other measurements, but also to recreate deconvolution results. The screenshot shows a few entries of a reference list based on a measurement in *C. glutamicum*.

the intensity of the same peak in the reference spectrum or zero. Two peaks in the spectra are considered the same, if their m/z values are within the user-defined m/z error ε_{mz}. A new annotation will be created for the best matching spectrum that achieves a cosine score above a user-defined threshold.

The user is able to add new spectra to (or delete them from) a reference list using the web interface. Fig. 5.9 provides a section of the reference list review page in ALLocator.

5.5 Data curation

LC-ESI-MS spectra are often convoluted to an extent that unambiguous solutions cannot be found by any tool yet. This is, because often multiple valid pseudo spectra are formed which share certain ions. Most deconvolution tools apply heuristics to return a most probable solution to this problem, generating datasets where any peak will only be part of one pseudo spectrum. These results thus do not reflect the ambiguity of the raw data and the user is not able to properly revise or curate the pseudo spectra. Eventually, erroneously allocated pseudo spectra can feign or hide metabolites, endanger a successful annotation in a measured sample and finally falsify scientific assumptions based on the experiment.

The data model of ALLocator allows to reflect the original ambiguity by setting monoisotopic peaks and pseudo spectra into a many-to-many relationship. On the one hand, this makes sure that deconvolution is not an irrevertible black box decision, on the other, it shifts a lot of responsibility to the user. Thus it is mandatory to provide the user with means to manually resolve conflicting spectra in an easy and transparent way. The ALLocator web platform puts its users into full control of all pseudo spectra, peaks, and metabolite annotations.

At first of course, the user needs to be informed about the current deconvolution results. A sortable and searchable table presents all putative molecules (i.e. pseudo spectra) and serves as the entry point to each individual spectrum (see Fig. 5.4). The 'Pseudo Spectrum' page of L-glutamate is given as and example in Fig. 5.10. A 'Pseudo Spectrum' page consists of three to four tabs, of which the tabs 'Metabolites', 'MassBank' (on demand), and 'Mass Decomposition' have been presented before in the Figs. 5.5, 5.6, and 5.8 respectively. The first tab, 'Pseudo Spectrum', provides a table comprising information on all ^{12}C monoisotopic peaks and the complete pseudo spectrum containing the same peaks as well as ^{13}C peaks and all isotopologues for both.

The sortable pseudo spectrum table offers seven columns: The m/z value, the RT in seconds, the ion annotation for this peak in this spectrum, the charge of the ion, the correlation of the extracted ion chromatogram (EIC) of the ion to the main adduct, the 'also starring in' column, and an additional column with a delete but-

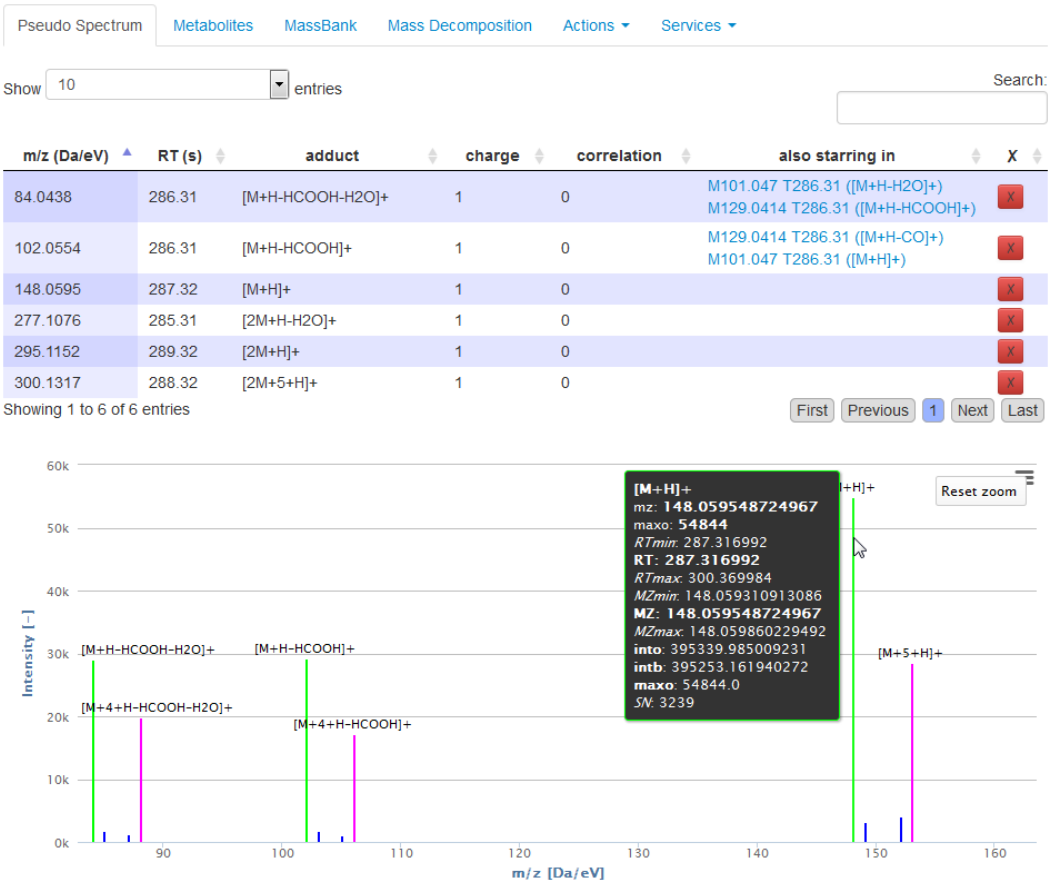

Figure 5.10: The screenshot shows the spectrum view of L-glutamate in a positive-mode measurement in a U-^{13}C SIL experiment. The view provides a table of all adducts and fragments contributing to the spectrum, along with the m/z-values and retention times of the ^{12}C isotopologue, the respective adduct specifications and charges. The 'correlation' column displays the correlation of the EICs of all ions to the main adduct ([M+H]$^+$ in this case). The column named 'also starring in' provides direct links to other pseudo spectra, to which the respective peak is associated with another adduct specification. The interactive visualization of the spectrum can be zoomed, provides additional information on peaks on mouse over, and allows to modify each peak association with a context menu.

ton which will remove the peak from the spectrum (after confirmation in a modal dialog). The 'also starring in' column is one of the unique features in ALLocator, which comes with the many-to-many relationship of peaks and pseudo spectra: Let L-glutamate be M_{147} and call the upper two adducts in the table $[M_{147}+H-HCOOH]^+$ and $[M_{147}+H-HCOOH-H_2O]^+$. These two fragment ions may as well be interpreted as $[M_{101}+H]^+$ and $[M_{101}+H-H_2O]^+$ of a second pseudo spectrum M_{101} or as $[M_{129}+H-CO]^+$ and $[M_{129}+H-HCOOH]^+$ of a third pseudo spectrum M_{129}. The latter spectrum is even supported by an additional peak $[M_{129}+H]^+$ which is not part of the putative L-glutamate spectrum. The 'also starring in' column informs the user about these conflicting interpretations and allows to quickly navigate to all pseudo spectra a certain peak is a potential part of. The different means to solve these conflicts are described below. Selecting an ion in the table will highlight the corresponding peak in the spectrum below.

The interactive spectrum shows all peaks that are part of the pseudo spectrum in the following color scheme: Monoisotopic ^{12}C peaks are green, monoisotopic ^{13}C peaks are purple, all isotopologues are blue (, and peaks that correlate well to the main adduct but are not part of the pseudo spectrum are gray.) The spectrum allows to zoom in both axes, m/z value and intensity. Hovering a peak with the cursor will trigger a tool tip to pop up, which contains additional detailed information on the peak as determined by XCMS (Smith *et al.*, 2006). Clicking any peak will open a peak-aware context menu. The here presented actions are tailored to the selected peak. A monoisotopic ^{12}C peaks can be removed from the pseudo spectrum, its annotation (i.e. ion specification) can be edited, or it can be used to create a new pseudo spectrum on its basis. A U-^{13}C peak can be marked as a not labeled peak, it can be converted to a regular isotope of its ^{12}C counterpart, or it can be used to create a new pseudo spectrum on its basis as a ^{12}C main adduct. Isotopologues can get their isotope status revoked.

While the above presented actions are theoretically sufficient to solve all conflicting peak-to-spectrum associations, there is a more convenient way. The function 'claim' in the 'Actions' drop down menu of the page removes all peaks associated to the current pseudo spectrum from all other pseudo spectra. After claiming all peaks, the pseudo spectrum will not be part of any conflict anymore. Other actions provided through this drop down menu are the deletion of a pseudo spectrum, adding this pseudo spectrum to a reference list or matching it against a reference list (cmp. subsection 5.4.5), loading the extracted ion chromatograms (EICs) of all monoisotopic peaks, and loading peaks that correlate with the main adduct peak of the current spectrum.

The user is warned, if a spectrum involving multiple ^{13}C peaks contains suspicious information in any of the two following ways:

The quotients $Q_{12:13;i} = I_{12;i}/I_{13;i}$ of n $^{12}C/^{13}C$ peak pair intensities have an co-

efficient of variance $c_v > 0.3$. Theoretically, the quotients $Q_{12:13;i}$ should be equal for all n ion species in a pseudo spectrum. If this is not the case, this is an indicator that the pseudo spectrum contains ions from different compounds.

The labeling information is inconsistent in such a way that the m/z distances $\Delta_{m/z}$ between ^{12}C and ^{13}C peaks do not reflect the annotated neutral losses: If in a given pseudo spectrum $\Delta_{m/z,a}$ of an [M+H]$^+$ ion is 12, $\Delta_{m/z,b}$ of an [M+H-$C_6H_{12}O_6$]$^+$ must be 6, as six of twelve carbon atoms have been lost. If this is not the case, the labeling information is marked as inconsistent.

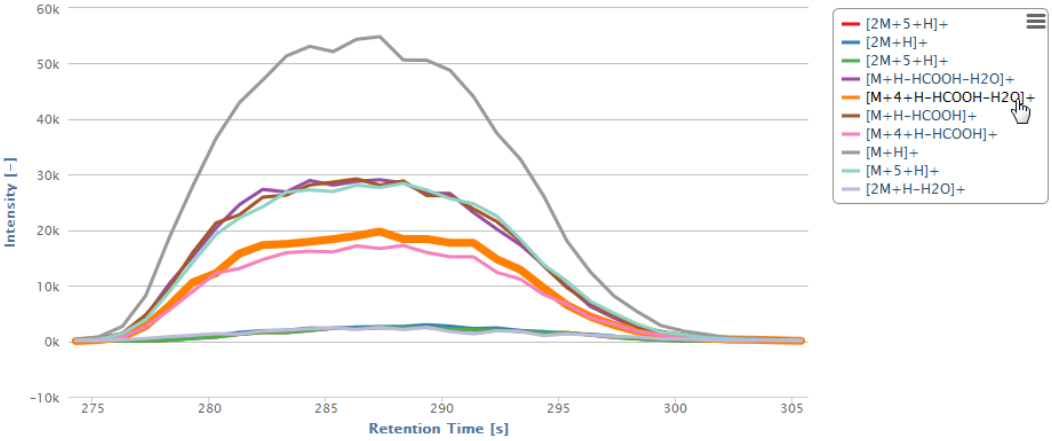

Figure 5.11: The EICs of all monoisotopic peaks in a spectrum can be loaded into the spectrum view on demand. It contains both monoisotopic ^{12}C and ^{13}C peaks. The visualization can be zoomed. The legend can be used to either highlight or hide EICs of certain peaks.

The EICs for the masses of all monoisotopic peaks are extracted from the raw data on the fly, using the XCMS (Smith *et al.*, 2006) library. These EICs are limited to the union of retention time ranges of all these peaks, such that all peaks of the pseudo spectrum are entirely covered. Fig. 5.11 shows the EICs of the L-glutamate pseudo spectrum mentioned above. It helps to assess the quality of the chromatogram and allows to easily spot peaks that have been assigned to the pseudo spectrum erroneously.

The user can also load other peaks into the pseudo spectrum which have not yet been associated to it, but correlate with the main adduct ion better than a user-defined threshold. Two options exist for this action: either to consider all peaks in the peak table, or to consider only peaks that have not yet been associated to any other pseudo spectrum. After loading correlating peaks, they will be added to the

pseudo spectrum in a gray color. Peaks that are associated to other pseudo spectra are displayed in a greenish gray to create awareness for the potential conflict. Correlating peaks can be associated to the pseudo spectrum in various ways using the peak-aware context menu: as a new adduct or fragment ion, as a U-^{13}C species of any ion, or as an isotopologue of any ion. They can also be used as a basis to create a new pseudo spectrum.

5.6 Summary

Correct metabolite identification in LC-ESI-MS datasets heavily relies on expert knowledge and cannot be done automatically per se. Due to this, metabolite identification is a major bottleneck in untargeted metabolomics experiments. In addition, stable isotope labeling was reported to greatly facilitate this process. ALLocator constitutes a novel powerful web platform for the semi-automatic annotation of peaks in LC-MS chromatograms and an interface that supports manual improvement of metabolite annotation with interactive tables and visualizations. A central cornerstone of this platform is the ALLocatorSD pipeline for the automatic assembly of pseudo spectra. As a major improvement compared to previously existing software, this new algorithm is capable of dealing with U-^{13}C labeling experiments, enabling not only relative quantitation, but also automatic annotation of fragments resulting from large neutral losses. For the subsequent manual revision and correction of automatic annotation results, the user benefits from the integration of the platform with public metabolite and mass spectral databases (KEGG (Ogata *et al.*, 1999; Kanehisa and Goto, 2000), ChemSpider (Pence and Williams, 2010), MassBank (Horai *et al.*, 2010)) and new powerful tools, as for example the spectrum-aware mass decomposition. The possibility to create, share and query user-defined reference lists is an important feature that ensures transferability of once made annotation efforts to other chromatograms and experiments. The system contributes to the metabolomics software landscape by extending the bioinformatics coverage of analytical technologies. By supporting LC-ESI-MS data and especially U-^{13}C SIL it complements the community of metabolomics online platforms, until now constituted by platforms like MeltDB 2.0 (Neuweger *et al.*, 2008; Kessler *et al.*, 2013), XCMSOnline (Tautenhahn *et al.*, 2012), and MetaboAnalyst (Xia *et al.*, 2009).

6 STUDY: Amino-acid profiling in *C. glutamicum* strains

In this study the ALLocator web platform was applied for the identification and relative quantitation of abundant metabolites in hydrophilic extracts of the *Corynebacterium glutamicum* type strain ATCC 13032 and the L-arginine producing (canavanine resistant) strain *C. glutamicum* ATCC 21831 (Nakayama and Yoshida, 1974). Four biological replicates were prepared for both strains and a U-^{13}C-labeled bacterial extract was used as internal standard. Cultivation of *C. glutamicum* strains, sampling and LC-MS analysis were carried out as described previously by Petri *et al.* (2013). Detailed mass spectrometer settings are listed in Tab. 6.1. Experimental raw data and protocols are publicly available (study identifier: MTBLS128[1]) through the MetaboLights repository (Salek *et al.*, 2013b). All chromatograms were uploaded to ALLocator and organized in a single experiment. Peak detection was performed using XCMS and resulted in the detection of approximately 1400 to 1500 peaks for each chromatogram (for XCMS parameter settings see Tab. 6.2). Subsequently, the ALLocatorSD algorithm was started to associate isotopologues and to generate pseudo spectra based on XCMS peak tables (for ALLocatorSD parameter settings see Tables 6.3 and 6.4). The molecule list view was then used for manual revision of peak annotations. At first pseudo spectra with a high number of peaks and those containing ^{13}C labeled peaks were reviewed. In addition, substrates and intermediates of the L-arginine biosynthesis pathway were specifically searched for. The complete procedure shall be demonstrated by the identification of glutamic acid, the most prominent metabolite and initial substrate for arginine biosynthesis in *C. glutamicum*. Using the search toolbar, the peak list was filtered to solely display metabolites with annotations containing the term glutamate. The list included a pseudo spectrum (M147.052T287.92) with six unlabeled peaks of a putative metabolite with a calculated neutral monoisotopic mass of 147.052 Da and a retention time of 288 seconds as depicted in Fig. 6.1. The neutral mass with a mass deviation of 0.01 Da matches 10 entries listed in the KEGG database, all for the same molecular formula $C_5H_9NO_4$ (see Tab. 6.5).

Mass decomposition for 147.052 Da was performed with all available filters ac-

[1]Study online available at: https://www.ebi.ac.uk/metabolights/MTBLS128

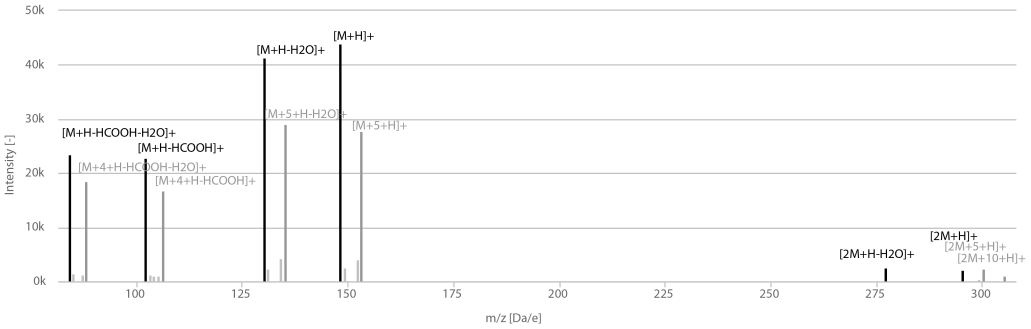

Figure 6.1: Automatically allocated pseudo spectrum M147.052T287.92 (glutamic acid), featuring ^{12}C monoisotopic peaks (black), ^{13}C monoisotopic peaks (medium gray) and associated heteroisotopic peaks (light gray). The Figure was exported from the ALLocator software and adapted for print.

tivated and finally resulted in only the single formula $C_5H_9NO_4$. This is indeed the molecular formula of L-glutamate, but also of all other nine metabolites listed in Tab. 6.5. A pseudo fragment spectrum was queried against the MassBank database. The best retrieved hit was a spectrum of glutamic acid (Glutamic acid; LC-ESI-QTOF; MS2; CE:15 eV; [M+H]+; MassBank: PB000462) with a score value of above 0.98. In fact, the list of fragment peaks in the pseudo spectrum was identical to that of the MS/MS spectrum of glutamic acid. All automatically annotated ^{13}C-labeled peaks and thereby inferred numbers of carbon atoms were consistent with the annotation of neutral losses, which was initially performed only on the basis of m/z differences. Intensity ratios for all ^{12}C monoisotopic peaks to their fully ^{13}C-labeled counterparts were similar. One labeled peak (m/z 300.1309) was automatically associated to the [2M+H]$^+$ adduct in a distance of +5 Da and annotated as [2M+5+H]$^+$. This peak most likely represented an adduct consisting of one unlabeled and one fully ^{13}C-labeled isotopologue. After searching for additional correlating orphan peaks, a peak was identified (m/z 305.1449) representing [2M+10+H]$^+$, the adduct of two fully ^{13}C-labeled glutamic acid molecules. This peak was manually added to the pseudo spectrum using the context menu (see pseudo spectrum in Fig. 6.1). All available information taken together enabled a reliable identification of glutamate, although no distinction between the L- and D-enantiomer was possible. A subset of peaks that are associated to a large pseudo spectrum can sometimes be added to an additional pseudo spectrum for another putative mass. This tends to happen when multiple consecutive small neutral losses occur. This shall be demonstrated again using the pseudo spectrum of glu-

tamate (147.0532 Da). Here, three of the peaks were annotated as $[M+H-H_2O]^+$, $[M+H-HCOOH]^+$, and $[M+H-HCOOH-H_2O]^+$. The same peaks were also assembled to the pseudo spectrum M129.04T287.92 and annotated as $[M+H]^+$, $[M+H-CO]^+$ and $[M+H-HCOOH]^+$, respectively. The putative neutral monoisotopic mass of this second pseudo spectrum (129.04 Da) matched for example 4-oxoproline in the KEGG database. As both pseudo spectra are formally correct when regarded separately and peak correlations can be very good for different coeluting compounds, this ambiguity cannot be solved reliably without manual revision. Thus, it is one of the main goals of the manual editing process to eliminate multiple annotations of such peaks. For this purpose the ALLocator function 'claim peaks' was used, which in this case deleted the mentioned peaks from all pseudo spectra except that of glutamate. This is an important advantage over editing annotations in a spread sheet, because it ensures data integrity and the concise visualizations help keeping the overview.

6.1 Annotation of large neutral losses allows identification of (γ−)glutamyl dipeptides

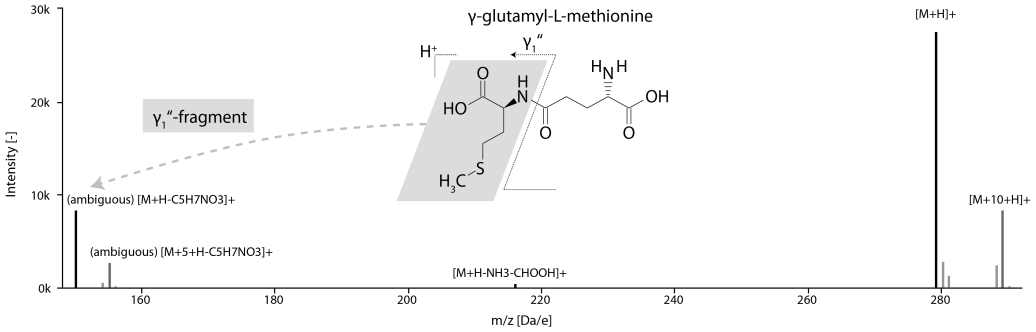

Figure 6.2: Pseudo spectrum and molecular structure of (γ−)glutamyl-L-methionine. The γ_1''-fragment (on the very left) could be annotated thanks to the additional information provided by the ^{13}C monoisotopic peaks; black: ^{12}C monoisotopic peaks; medium gray: ^{13}C monoisotopic peaks; light gray: associated heteroisotopic peaks; the '(ambiguous)' tag informs the user that this neutral loss was calculated by mass decomposition and filtered for the correct number of carbon atoms, but still multiple molecular formulas might explain the present mass difference.

Amongst the metabolites with the most prominent peaks in both strains sev-

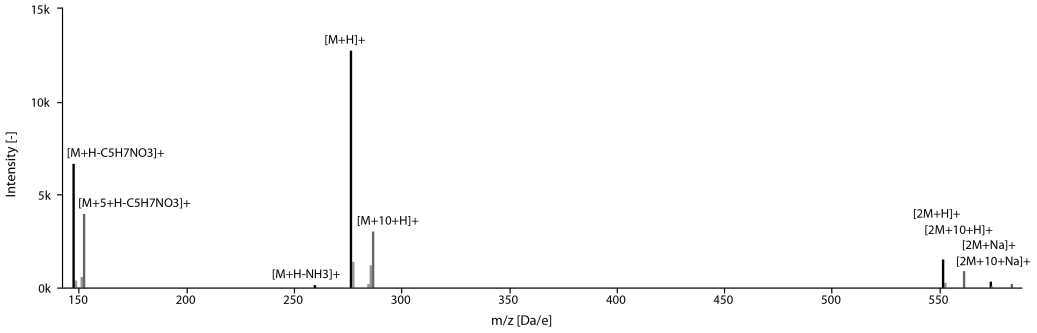

Figure 6.3: Pseudo spectrum of (γ-)glutamyl-glutamine. Black: ^{12}C monoisotopic peaks; medium gray: ^{13}C monoisotopic peaks; light gray: associated heteroisotopic peaks.

eral dipeptides were identified. The calculated monoisotopic masses all matched those of at least two different peptides, containing a glutamyl residue at the N- or the C-terminal end. On the basis of the calculated mass alone it was not possible to distinguish between the isobaric compounds, but positional information could be inferred from the generated pseudo spectra. These included peaks for the respective γ_1''-fragment of the peptide (Fig. 6.2), showing that all dipeptides had an N-terminal glutamyl residue (see Figs. 6.2 to 6.5).

The automatic annotation of the γ_1''-fragments was possible through the unique ability of ALLocatorSD to deal with ^{13}C-labeling experiments. These uncommon fragments are not included in the list of small neutral losses, but could be annotated in the sixth step of the ALLocatorSD pipeline (see section 5.3). This aided the annotation of the dipeptides as glutamyl-methionine, glutamyl-valine, glutamyl-(iso)leucine and glutamyl-glutamine. In case of glutamyl-glutamine the γ_1''-fragment was not assigned by ALLocatorSD. The tool 'find correlating peaks' was used with a lowered correlation coefficient threshold of 0.75. The peaks (m/z 147 and m/z 152) representing the expected γ_1''-fragment and its fully labeled ^{13}C isotopologue were present and added to the pseudo spectrum (see Fig. 6.3). Checking the extracted ion chromatograms (EICs), a different peak shape for these m/z values and a slightly higher retention time compared to the other peaks of the pseudo spectrum was observed. Additionally, the intensity of the fully ^{13}C-labeled peak compared to the ^{12}C monoisotopic peak was higher than for all the other peak pairs. All these differences could be referred to the coelution of free glutamine, which was checked by the analysis of L-glutamine standard. Previously, γ-glutamyl-L-glutamine, γ-glutamyl-L-valine, γ-glutamyl-L-leucine and γ-glutamyl-L-glutamate have been isolated from *C. glutamicum* fermentation broths, but the physiological

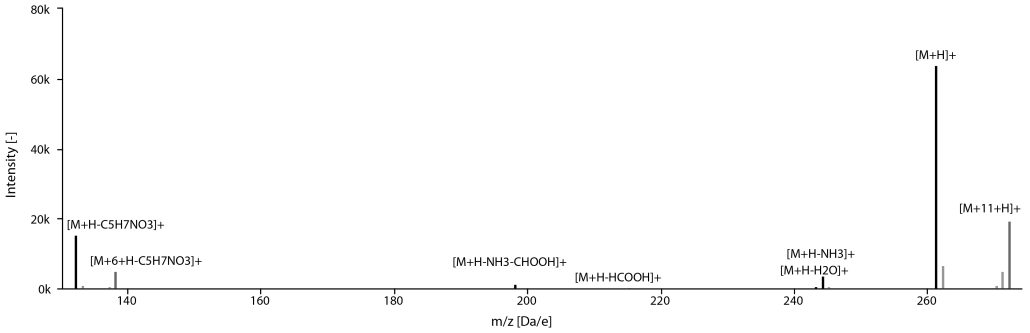

Figure 6.4: Pseudo spectrum of (γ-)glutamyl-leucine. Black: ^{12}C monoisotopic peaks; medium gray: ^{13}C monoisotopic peaks; light gray: associated heteroisotopic peaks.

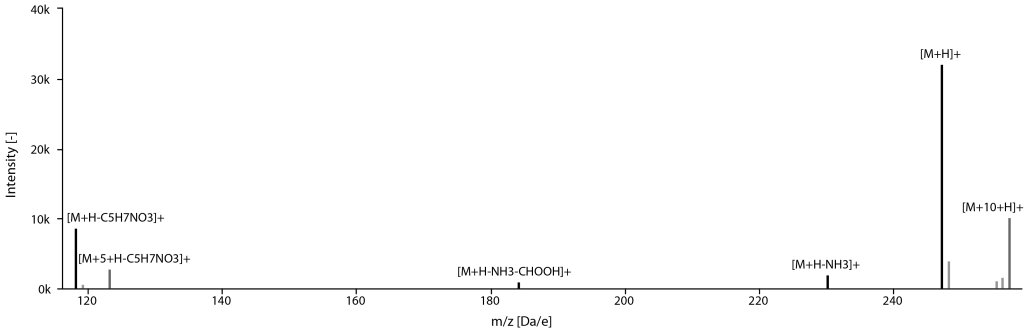

Figure 6.5: Pseudo spectrum of (γ-)glutamyl-valine. Black: ^{12}C monoisotopic peaks; medium gray: ^{13}C monoisotopic peaks; light gray: associated heteroisotopic peaks.

role of these metabolites stayed elusive (Vitali *et al.*, 1965; Hasegawa and Matsubara, 1978). Although the presence of $[M+H\text{-}NH_3]^+$ and absence of $[M+H\text{-}H_2O]^+$ ions in the spectra of the before mentioned peptides were an indication for γ-linkages (Harrison, 2003), it was not possible to readily distinguish between dipeptides with α- or γ-linkages. Kessler *et al.* (2014) is the first report on the synthesis of (γ-)glutamyl-methionine by *C. glutamicum*, but amongst other γ-glutamyl dipeptides it was detected earlier for example in samples of *Synecococcus sp.* PCC 7002 by an untargeted metabolomics approach (Baran *et al.*, 2010, 2013).

In order to safe the manual annotation effort and to transfer it to all the other chromatograms in the experiment, the curated pseudo spectra with confirmed metabolite annotations were stored in a reference list using the tool 'create reference spectra'. This reference list was later used to automatically detect, assemble and annotate similar pseudo spectra in all the other chromatograms of this experiment using the function 'apply reference list'.

6.2 Data export and relative quantitation of arginine biosynthesis intermediates

The identification of bottlenecks by the detection of accumulating pathway intermediates in large libraries of strains is an integral part of modern metabolic engineering strategies and biotechnology (Pitera *et al.*, 2007). To demonstrate functionalities for export of data and relative quantitation of metabolites, it was obvious to compare the relative abundances of metabolites of the arginine biosynthesis pathway, since *C. glutamicum* ATCC 21831 is an L-arginine producing strain. Here it was possible to identify the substrates L-glutamate and L-glutamine, the intermediates N-L-acetylglutamate, L-citrulline and N-L-argininosuccinate, as well as the endproduct L-arginine. For each confirmed metabolite, peak intensities and areas were automatically normalized to internal standard and biomass, and exported to an xls document. Relative quantitation between sample groups and statistics were performed in a spreadsheet (see Tab. 6.6). Metabolites mentioned in the following were quantified using the peak areas of the respective $[M+H]^+$ ions, and all their abundances were significantly different between strains. The significance was determined by Student's t-test and multiple testing errors were corrected using the method of Benjamini and Hochberg (Benjamini and Hochberg, 1995). The concentration of the initial substrate L-glutamate was lower in the arginine producer than in ATCC 13032 (fold-change 0.23). The intermediate N-acetylglutamate was detectable in all samples, but the peaks of the ^{13}C-labeled internal standard were below the detection limit, so that no relative quantitation could be performed. As expected, the L-arginine pool was higher (fold-change 12.26) in *C. glutamicum* ATCC 21831 compared to the type strain. But in addi-

tion, accumulation of N-L-argininosuccinate (fold-change 34.05) and L-citrulline (fold-change 1.9) could be observed, indicating a bottleneck in the last step of arginine production, the conversion of N-L-argininosuccinate to arginine and fumarate. This is in good accordance with a study by Park *et al.* (2014), in which the strain ATCC 21831 (AR0) was used in a systems metabolic engineering approach. Here, authors debottlenecked the last two reactions of the arginine biosynthesis in the derived strain AR6 by replacing the native promoter of the argGH operon with a stronger one.

6.3 Discussion

This application study demonstrated the applicability of the ALLocator web platform on complex biological samples and the software was used to annotate and relatively quantify intermediates of the L-arginine biosynthesis in two strains of *C. glutamicum*. Analyzing the data specifically with regard to arginine biosynthesis, the last step of the pathway was identified as a bottleneck in L-arginine production with strain ATCC 21831. In an untargeted manner γ-glutamyl-methionine, a previously unknown metabolite of *C. glutamicum*, has been identified. By providing tools for widely automated identification, quantitation and exploration of LC-ESI-MS data, ALLocator is well suited for the processing of LC-ESI-MS datasets in the fields of systems biology and biotechnology.

Table 6.1: Parameters for microTOF control in full scan MS mode.

Mode			
Scan Mode	MS	Ion Polarity	Positive
Mass Range	50-1000 m/z	Rolling Average	off
Spectra Acquisition	Save Spectra	Absolute Threshold	10
Include Profile Spectra	Always	Peak Summation Width	5 pts
Focus	Inactive	Acquisition Rate	1.0 Hz
Source			
Endplate Offset	-500 V	Dry Gas	8.0 L/min
Capillary	-2500 V	Dry Temp	180 °C
Nebulizer	3.0 bar		
Transfer			
Funnel 1 RF	180.0 Vpp	ISCID Energy	0.0 eV
Funnel 2 RF	200.0 Vpp	Hexapole RF	100.0 Vpp
Quadrupole			
Ion Energy	5.0 eV	Low Mass	100.0 m/z
Funnel 2 RF	200.0 Vpp	Hexapole RF	100.0 Vpp
Collision Cell			
Collision Energy	10.0 eV	Collision RF	150.0 Vpp
Transfer Time	70.0 μs	Pre Pulse Storage	7.0 μs

Table 6.2: Parameters for XCMS processing as set within the ALLocator GUI.

Parameter	Value
ppm	30
peakwidth_min	10
peakwidth_max	60
noise	0
snthresh	10

Table 6.3: Parameters for ALLocatorSD processing as set within the ALLocator GUI.

Parameter	Value
rtBegin [s]	73
rtEnd [s]	511
epsilonMZ [Da/e]	0.005
epsilonRT [s]	5.0
minCorrelation	0.75
minIntensity	0
labeled13C	true
omit100DaMassDecomps	true

Table 6.4: Adduct settings used for spectra deconvolution in ALLocator.

primary	seed	common	excluded	name
x				[M+H]+
		x		[M+Na]+
		x		[M+K]+
		x		[M+2Na-H]+
		x		[M+2K-H]+
		x		[M+H-H2O]+
		x		[M+H-H2O-H2O]+
		x		[M+H-H2O-H2O-H2O]+
		x		[M+H-HCOOH]+
		x		[M+H-HCOOH-H2O]+
		x		[M+H-C2H4O2]+
		x		[M+H-NH3]+
		x		[M+H-CO]+
		x		[M+H-CO2]+
		x		[M+H-C6H10O5]+
		x		[M+H-C6H12O6]+
		x		[M+H-C6H12O6-H2O]+
		x		[M+H-C2H2O]+
		x		[M+H-C2H5O2]+
		x		[M+H-NH3-CO2]+
		x		[M+H-NH3-CHOOH]+
		x		[M+H-CH5N3]+
		x		[M+H-CH4N2O]+
		x		[M+2H]2+
		x		[M+2Na]2+

Table 6.5: The Pseudo Spectrum for M147.052T287.92 (glutamic acid) matched 10 KEGG Compound entries by the putative neutral mass 147.052 ± 0.01 Da, which all feature the same molecular formula $C_5H_9NO_4$.

KEGG Compound Name	KEGG ID
L-Glutamate	C00025
D-Glutamate	C00217
Glutamate	C00302
O-Acetyl-L-serine	C00979
L-threo-3-Methylaspartate	C03618
N-(Carboxymethyl)-D-alanine	C03790
Isoglutamate	C05574
L-4-Hydroxyglutamate semialdehyde	C05938
2-Oxo-4-hydroxy-5-aminovalerate	C05941
N-Methyl-D-aspartic acid	C12269

Table 6.6: Normalized peak areas of pseudo-molecular $[M+H]^+$ ions of different metabolites.

	ATCC 21831		ATCC 13032				
Metabolite	Average	SDEV	Average	SDEV	Ratio	p-value	q*
L-Glutamate	0.20589	0.05522	0.89333	0.04974	0.23*	1.8E-06	0.0125
L-Citrulline	0.25210	0.06581	0.13051	0.01081	1.93*	0.03260	0.0500
N-(L-Arginino)succinate	0.31388	0.09093	0.00922	0.00137	34.05	0.00677	0.0375
L-Arginine	0.69132	0.10437	0.05638	0.00224	12.26	0.00119	0.0250

7 MeltDB 2.0

Compared to other molecular levels or -omics methods, metabolomics is challenging in its high degree of interdisciplinarity, interlinking experts from research fields as diverse as engineering, physics, chemistry and biology and from cheminformatics over bioinformatics to statistics, data mining and finally visualization.

Both sample acquisition and subsequent analysis are automated in high-throughput instruments which has continuously posed challenges on the systematic storage and computational processing of the gathered experimental datasets, starting in the early 2000s. The increasing number and quality of measurements not only raised the generated data volume but also allowed to address more complex biological questions within conducted experiments. To comprehensively address these demands bioinformatics internet applications were developed. MeltDB, 'a software platform for the analysis and integration of data from metabolomics experiments', has been published by Neuweger *et al.* (2008). Xia *et al.* (2009) released MetaboAnalyst, 'a comprehensive tool suite for metabolomic data analysis'. Carroll *et al.* (2010) published the MetabolomeExpress web server as 'a public place to process, interpret and share GC/MS metabolomics datasets.'

Since around 2008 we have observed that the requirements to comprehensive metabolomics software platforms have changed: The general growth of the field of metabolomics and the increasing number of collaborations diversified the user community of researchers and their individual scientific goals. It is obvious that the success of a metabolomics study depends on an efficient and effective collaboration of this interdisciplinary research community. Thus not only the availability and sharing of the data is important, but also special functions have to be significantly extended with specific features to consider all researcher's demands and perspectives. In addition, the ever increasing throughput and the constant lack of time makes it immensely important that automated pre-processing methods are reliable and that analyses and manual intervention are fast and easy. And since metabolomics approaches are applied to more and more scientific objectives, a powerful set of statistical methods is the natural next requirement, ranging from hypothesis-driven statistical tests to less specified and untargeted data mining methods, such as clustering and dimension reduction. Finally, the wealth of generated data poses a necessity for interactive data exploration and curation (IDEC) tools.

To tackle these new challenges systematically, a next generation of bioinformat-

ics tools needed to be developed, covering all of the above mentioned aspects of metabolome data analysis, ranging from preprocessing raw data to data integration and finally the derivation of biological knowledge. These steps were comprised above in Fig. 4.1 as *raw data acquisition*, *data preprocessing*, *data integration*, *data analysis*, and *interactive data exploration and curation*. During the stages of that process one can also identify four successive data categories that represent the different levels of data classification and annotation as well as different levels of abstraction. These describe the state of the data before and after each major step of the computational metabolomics workflow: First, *raw data* (RD), stored and organized in meaningful groups after acquisition, builds the basis. Then, *pre-processed data* (PD) is computed, where peaks and their quantities have been detected. It follows *integrated data* (ID), where peaks that putatively originate from the same compound are consistently annotated over chromatograms of an experiment and thus become comparable. Last, *derivative data* (DD) is achieved by statistical analyses of metabolite quantities in an experiment and then visualized to allow effective exploration and to draw conclusions.

The following sections present MeltDB 2.0, which offers novel tools to challenge the rising wealth of data quality and quantity and support the analysis of all four categories RD, PD, ID, and DD and includes a multitude of updates. Sophisticated preprocessing methods underpin the reliability of automatically created annotations. At the same time, straight forward tools for manual peak annotation simplify the curation even of large experiments. To help answering questions of different scientific objectives, a rich set of statistical analyses and data mining tools is made available. To finally nail down the quintessence of an experiments outcome, data exploration is supported by interactive and telling information visualizations.

7.1 System design and implementation

The first version of the MeltDB software platform, a 3-tiered web application and database server published in 2008 (Neuweger *et al.*, 2008), provides means for the standardization, systematic storage and analysis of GC-MS metabolomics experiments. Within a powerful project and user management, raw chromatograms of various file formats can be uploaded and organized into chromatogram groups (e.g. replicates, factor levels) and experiments. A flexible processing pipeline allows to find, quantify and identify peaks in the raw chromatograms. Subsequently, a set of statistical tools and visualizations can be applied to analyze the gathered data tables. This fast growing, free online platform today hosts more than 25 distinct projects conducted by more than 150 registered users from around the world. More than 17,000 chromatograms have been uploaded and analyzed yet.

7.1.1 System integration

The MeltDB 2.0 web system serves as a platform including a number of tools to provide their features to the user in an integrated fashion. The paragraphs of this subsection outline the technical aspects and software design that build the foundation of this platform.

Platform

The core value of MeltDB as a web platform is its carefully designed object relational database model, which combines metabolomics experimental data, metadata, and the parameterizations and results of preprocessing tools. The model design was created using the O2DBI software (unpublished). O2DBI automatically generates Perl code for data access object classes with *create*, *retrieve*, *update* and *delete* (CRUD) functionality, each one mapped to a specific class in the data model. These allow to focus on the development of business logic and features, rather then 'boilerplate' code. MeltDB is thus easy to extend and can serve as a platform for 'Software as a Service' (SaaS), providing a variety of features and tools of (GC-MS-based) metabolome informatics (Neuweger, 2009).

7.1.2 System design and data model

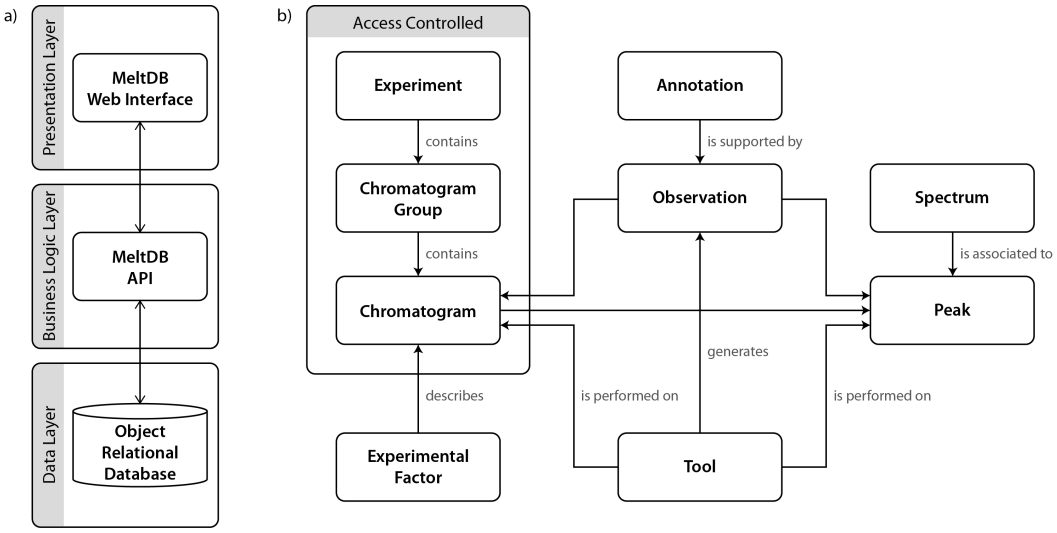

Figure 7.1: a) The layers of the MeltDB three tier architecture; b) schema of the main classes of the MeltDB data model and how they interact with each other. Reproduced from Neuweger (2009)

.

The MeltDB system design is built on a three tier architecture, in which one layer is reserved for each, data model, business logic, and the presentation (see Fig. 7.1a). A brief overview of the most important classes in the data model and their relations is given in Fig. 7.1b. In the business logic layer the MeltDB API provides *create*, *read*, *update*, and *delete* operations for all modeled classes. This core functionality is extended with more specific API where necessary. The MeltDB web interface finally constitutes the presentation layer, providing access to and visualizations of the data to the user (Neuweger *et al.*, 2008).

7.1.3 User management

As MeltDB is part of the CeBiTec bioinformatics software platform, projects, users and restricted access are managed by the General Project Management System (GPMS). The GPMS ensures data security through permissions imposed on individual tables of the (MySQL) database. As is depicted in Fig. 7.1b, access to experiments, chromatogram groups, and chromatograms is furthermore restricted. Here, additional permissions may be granted on a very fine granularity and are enforced through the business logic (Neuweger *et al.*, 2008).

7.2 General workflow and integrated features

A scheme of the general workflow, listing the available tools and visualizations for all steps from raw data to the IDEC of derivative data, is presented in Fig. 7.2. More information on preprocessing, profiling, data display and data mining in MeltDB 2.0 is given in the upcoming subsections.

7.2.1 Methods for data preprocessing

In computational metabolomics preprocessing is a critical step, as *integrated data* and *derivative data* build upon *preprocessed data*. In order to ensure a reliable data basis for statistical data exploration MeltDB 2.0 is equipped with a variety of algorithms for the early steps of experiment data analysis.

The growing list of preprocessing methods includes support for the centWave algorithm by Tautenhahn *et al.* (2008) for chromatographic peak detection which features a high sensitivity, and updates of the XCMS package (Smith *et al.*, 2006) for chromatogram alignment and profiling analyses. In addition, the ChromA (Hoffmann and Stoye, 2009) software adds to the list of supported chromatogram alignment tools. ChromA computes pairwise alignments of chromatograms without *a priori* knowledge, but is capable of optionally using previously matched or identified peaks as anchor points which speeds up the process.

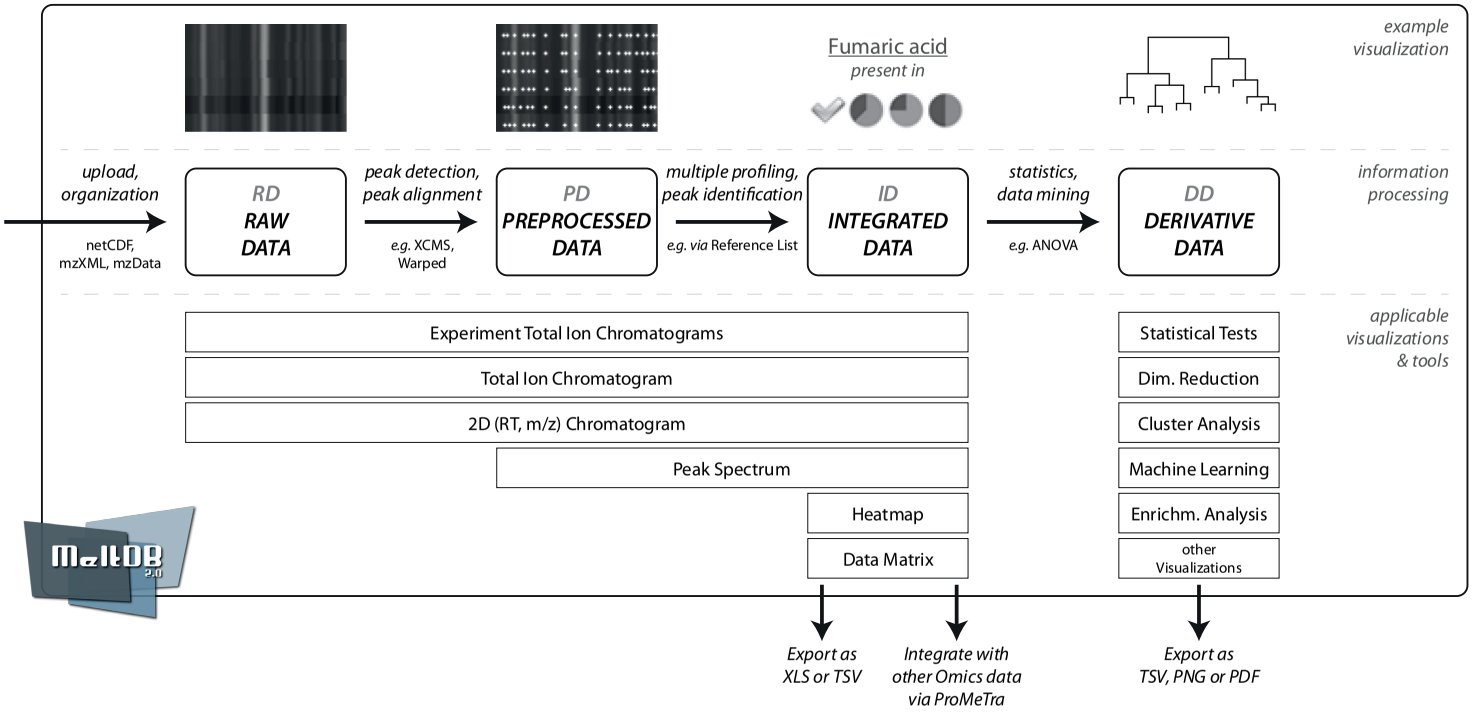

Figure 7.2: The overview shows the information processing in MeltDB 2.0 as well as visualizations and tools that are applicable to each level of data: RD, PD, ID and DD. Although different chromatogram viewers are available immediately after RD upload, heatmaps and data matrices can only be computed as soon as data have been integrated, i.e. there are peaks that are consistently named across chromatograms. To finally derive knowledge from the data, MeltDB 2.0 offers a versatile set of statistics and data-mining tools.

The automatic calculation of retention time indices in GC-MS measurements is furthermore aided by a web interface for manual index assignment. Peaks of added substances can be assigned with retention indices and will be used as anchors for interpolating other peaks retention indices (Ettre, 1994) which support subsequent peak identification (Kopka *et al.*, 2005). The detection of alkanes as retention markers can be automated.

Furthermore, peak identification itself is facilitated with a powerful feature: MeltDB 2.0 offers a *Reference list* tool to save peaks of measured reference substances as *Reference* in the MeltDB database (see 2.5.2 Spectra matching and cmp. 5.4.5 Custom reference lists). The stored data comprises retention indices, quantification masses and mass spectra of reference compounds. This helps to generate project specific databases that complement the GMD[1] (Kopka *et al.*, 2005) or the NIST standard reference database 1A[2]. The tool allows to aggregate *References* and to use their underlying mass spectra for efficient peak identification and comparison.

7.2.2 Profiling methods for data integration

To complete the first step of *data integration*, peaks in different chromatograms that derive from the same small molecule have to be named consistently and need to be associated to each other. Thus, support for GC-MS-based metabolite profiling experiments has been implemented. The focus for the profiling approach in MeltDB 2.0 is to combine the results from chemometrics approaches with further identification.

The generic MeltDB approach can be applied on netCDF, mzXML (Pedrioli *et al.*, 2004), and mzDATA (Orchard *et al.*, 2007) measurements from any supported analytical system. The novel tool registers peaks for non-targeted metabolite profiling based on multiple criteria. These are similarity of mass spectra, retention time difference and the existence of common EIC peaks above a given signal-to-noise threshold. It allows to annotate completely unknown peaks that are consistently detectable in several measurements of a metabolomics experiment. Thus, these potentially interesting peaks can be subjected to further statistical analyses in MeltDB and become accessible for profiling experiments, where the aim is in general to find differences in the metabolic composition of two or more sample groups.

Parameterizations of all pre-processing tools can be customized freely from within the web interface. Parameterizations are persisted project wide, so that other users from the same project - who typically use similar instrumental setups - can reuse them.

[1] http://gmd.mpimp-golm.mpg.de/
[2] http://www.nist.gov/srd/nist1a.cfm

Since pre-processing methods are usually too long running tasks as to complete during a web request or to even allow interactivity, these jobs are forwarded to the compute cluster of the CeBiTec bioinformatics platform and run decoupled from the web server.

As soon as pre-processing is completed at least to the point of profiling, data tables can be exported as XLS or TSV files. This allows to subsequently use the MeltDB 2.0 processing results in other, external programs of choice.

7.2.3 User interfaces for all levels of data abstraction

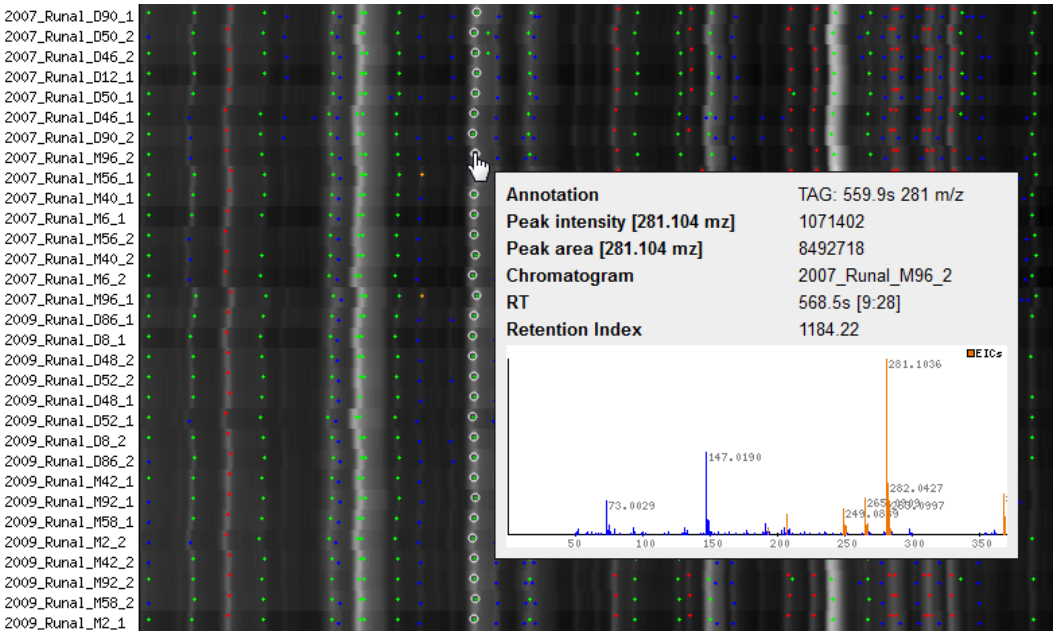

Figure 7.3: The *Experiment Total Ion Chromatogram* view depicts the total intensities over time for all chromatograms of an experiment. Detected peaks are marked with colored dots. Blue: Detected peak; Green: peak that was consistently detected throughout chromatograms; Red: Peak that was identified using any database. Hovering a peak highlights other peaks (*link-and-brush*) with the same annotation and description and shows a tooltip with detailed information. From this view, semi-automated tools for annotation can be accessed.

Visual inspection and the possibility for IDEC is important at all data abstraction levels, i.e. from *raw data* to *derivative data*. Intuitive and responsive data visualizations are required as well as intelligent tools that allow to solve common tasks,

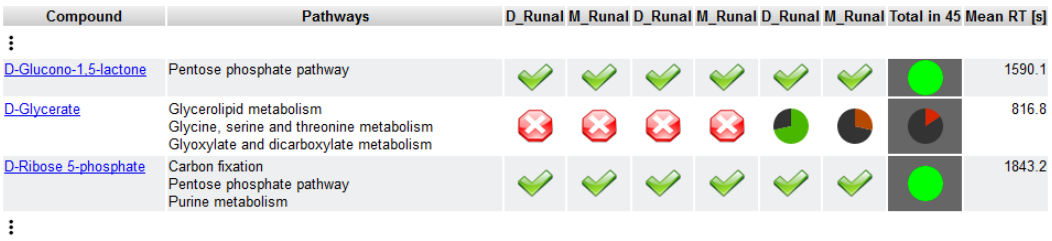

Figure 7.4: The *Show Absent/Present Compounds* view lists which compounds were identified in which chromatogram groups in how many chromatograms (pie charts). Checked fields tell that the respective compound was identified in all chromatograms of that group. The darker column in the right represents the entire experiment. This view allows to assess the completeness of compound annotations in an experiment.

like the inspection of processing results or the navigation from an experiments overview to the spectrum of a single peak, in a few steps.

Introducing *Asynchronous Javascripting and XML* (AJAX) to MeltDB 2.0 using the jQuery library[3] allows a new quality of interactive and dynamic data display[4] in terms of speed, responsiveness and user guidance. This was utilized to improve the *Experiment Total Ion Current* view (cmp. Fig. 7.3) with peak specific tooltips containing information about the associated compounds, latest annotations, and quantities. On demand, the complete peak object can be loaded and displayed inside the experiment TIC view, giving i.a. access to its spectrum, the complete list of annotations and observations.

One additional major improvement that benefits from the employment of AJAX is the dynamic manual annotation dialog. The interactive annotation functionality for whole experiments has been improved, so that aligned peaks with high mass spectral similarity can be annotated in parallel. The streamlined web interface helps to annotate peaks across chromatograms in a consistent manner, when researchers annotate manually, correct errors of automated annotation tools, or correct data that has been imported beforehand. This consistency is ensured by an autocompletion of compound names according to the KEGG database and is important for statistical analyses and comparison of results among different experiments.

The *Show Absent/Present Compounds* view (see Fig. 7.4) helps to assess the completeness of annotations, resulting from joint automated and manual annotation

[3] http://jquery.com

[4] The user can interact with visualizations to cause small changes to the data representation. These changes are performed in place without requiring a reload of the visualization.

Figure 7.5: The Flash® based heatmap visualization tabulates color-encoded relative abundances of metabolites in either chromatograms or chromatogram groups (shown). Abundances may be normalized on the entire experiment or per metabolite (shown). Rows and columns can be sorted freely or automatically by alphabetical order or according to the coefficient of variation. A gain factor may be set, to reveal differences between relatively small abundances.

efforts. Not only does it give a quick overview of compounds with a high recovery rate throughout all analyses, but it also provides first insights into differences between sample groups: The results are shown for the entire experiment and additionally split into the groups defined by the single factorial experiment design.

7.2.4 Statistics and data mining

In metabolomics, data analysis and mining can be driven by various intentions and will strive for *derivative data* with different purposes or try to aid data exploration and curation by highlighting the most relevant data subsets. One way to group these aims is to relate them to one of the following questions: *a)* What are the significant features of a sample group separating it from other sample groups? *b)* Do groups of samples form clusters according to their features and quantities? *c)* Can a sample be assigned to a class based on its spectral features? For each of

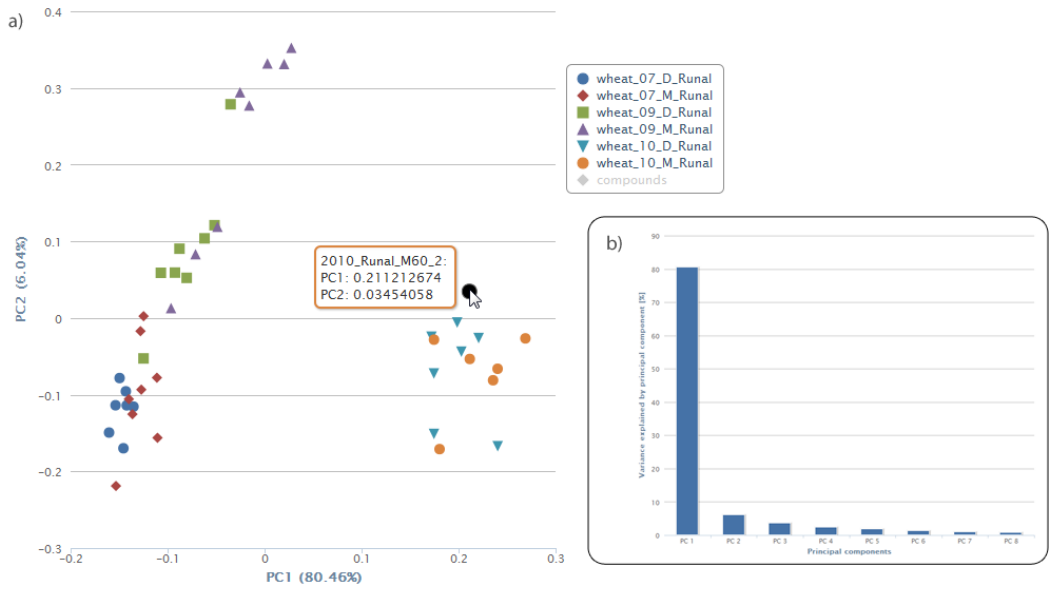

Figure 7.6: *(a)* The principal component analysis (PCA) is one of the most applied work horses in MeltDB 2.0. The visualization is implemented using Highcharts components. Single chromatogram groups can be shown and hidden on the fly. A zoom functionality is available, which is extremely useful in large experiments. Each data points chromatogram name and coordinates can be revealed via tooltip. Screenshot *(b)* shows the loadings plot of the same PCA.

these questions a variety of statistical analysis methods exists.

A binding to the R software (R Development Core Team, 2011) makes numerous statistical tools available from within the MeltDB software (Neuweger *et al.*, 2008). Its database objects are converted to R objects in a standardized manner. Information about chromatograms associations to chromatogram groups is pertained in the data representation. This integrated data conversion avoids the cumbersome process of converting data tables from proprietary software to a format a statistics software package can interpret and analyze. By default data is gathered from the database on the fly, but snapshots can be stored to speed up statistical analyses vastly, especially for large experiments.

Statistical analyses and data mining tools in MeltDB are accessed through a standardized parameterization and data selection form. Where appropriate, this basic form is extended with specific options and parameters. The basic form consists of a list of all chromatograms of the experiment from which the user may select. Additionally, features can be selected to be considered in the analysis. Features can

either be identified compounds or unidentified features that have been detected consistently among most chromatograms. A feature or compound can be chosen as reference to normalize to. When available, ribitol is preselected. One can select whether peak intensities or peak areas will be used for quantitation. Quantitations can be scaled linearly or logarithmically. Missing values can be handled in different ways.

To determine the significance and variances of features (see question *a*), the t-test of the Perl CPAN package *Statistics::TTest* (Juan, 2003) and *Statistics::KruskalWallis* (Lee, 2003) is offered as well as analysis of variance (ANOVA) using the *aov* method of the R statistical software (R Development Core Team, 2011) which fits a linear model. For all of these Bonferroni, Holm and Benjamini & Hochberg corrections are calculated (Holm, 1979; Benjamini and Hochberg, 1995). For each feature (metabolite) presented in the ANOVA and Kruskal Wallis test results a boxplot view and the extracted ion chromatograms of all samples can directly be accessed. Insight into different abundance levels and the coefficients of variation in samples and groups can be interactively achieved using the heat map visualization presented in Fig. 7.5.

In another view, the m-values (log-2 signal ratios) of features of all chromatogram groups of an experiment in reference to the same features in a user-selected chromatogram group are displayed tabularly. Volcano plots can be created plotting either the a-values (average log-2 signal values) or t-test values (as negative decadic logarithm of the p-value) against m-values. Variable importance estimation via the random forest algorithm from the *caret* R package (Kuhn *et al.*, 2011) can be applied to find differing features in groups. The metabolite set enrichment analysis published by Persicke *et al.* (2011) is another powerful tool in MeltDB for the identification of differentially regulated metabolic pathways.

Samples may aggregate to clusters according to their features quantities (see question *b*), regardless of the groups they nominally belong to. To visualize these clusters MeltDB provides the dimensionality reduction methods principal component analysis (PCA, *prcomp* method in R, cmp. Fig. 7.6), independent component analysis (ICA, fastICA package for R) and partial least squares discriminant analysis (PLS-DA, *caret* package for R) (R Development Core Team, 2011; Marchini *et al.*, 2010; Kuhn *et al.*, 2011). Hierarchical clustering allows to display dendrograms of chromatograms and is made available using the *hclust* method in R which can be applied with different linkage methods. The *heatmap* method of R is used to show false color maps of feature signals in chromatograms, sorting columns and rows according to the before-mentioned hierarchical clustering of feature signals and chromatograms, respectively. Here, data can be normalized for either chromatograms or features.

Whenever a dataset is subdivided in k groups (see question *c*), since samples were for instance taken from k sites or treated with k different protocols, another

suitable data mining strategy is to apply supervised machine learning methods to learn to approximate a relationship between metabolic profiles to the k categories (Hastie *et al.*, 2009). Such a classification can be helpful in the design of automated screening and identification processes or give insight into hidden links in small molecule patterns which are characteristic for a group k'. This propelled to extend MeltDB with the powerful R package *caret* (Kuhn *et al.*, 2011) of which the variable importance estimation has been mentioned above. Now, classification algorithms (support vector machine, random forest) can be trained with chromatogram groups representing c different classes $\{\omega_0, .., \omega_{c-1}\}$ and then be applied to other chromatograms of samples which have not yet been assigned to any class ω_i. For evaluation purposes, the user may opt to partition chromatograms into training and testing groups randomly. Additionally, MeltDB uses *caret* to compute and evaluate the classification performances of the algorithms random forest, k nearest neighbors, support vector machine, neural networks, and partial least squares, to estimate which classification algorithm performs best on a problem.

Generally, the computed results can be downloaded as TSV, XLS, PNG or PDF files in addition to the representation in the web browser.

7.3 Summary

MeltDB 2.0 was developed to comprehensively provide means to complete the entire process from *raw data* to *derivative data* within a software platform that supports researchers of diverse scientific backgrounds and fosters collaborations in complex metabolomics research projects. It was a further goal to make the final exploration of produced results and statistical outcomes effective and efficient. This has been achieved by improving the MeltDB tool set throughout all four stages *raw data*, *pre-processed data*, *integrated data* and *derivative data*. These recent developments leveraged MeltDB to an interactive rich internet application that allows to generate high-quality data sets and to dive deep into their analyses.

Since MeltDB was first published in 2008, a few other tools have been released that take a similar line. The MetaboAnalyst (2.0) web server offers a feature set similar to MeltDB, also supplying means to cover the pipeline from *raw data* to *derivative data*. MetaboAnalyst is merely made for a one-time, web service-like usage though, while MeltDB offers a project and user management that supports collaborative work, allows to manually refine and annotate processing results, and stores data for documentation purposes and to support larger and/or long-term projects.

The MetabolomeExpress web server is dedicated to making reviewed data sets publicly available. For that, it offers a fixed pipeline and a set of statistical tools which comprises a clearly smaller set of features, compared to MeltDB or Metabo-

Analyst.

The hierarchical data model of MeltDB 2.0 serves single-factorial experiment designs best. It is still possible to address even complex multi-factorial designs, but then a careful organization (sometimes multiple organizations) of chromatogram groups is necessary. The application example shown is a multi-factorial experiment differentiating wheat samples of different years and farming schemes. It is part of a study that additionally considers different cultivars as a third factor, which was also investigated in MeltDB 2.0.

Despite the advantages and opportunities of web platforms, this technology also has its drawbacks. The most critical aspect probably is the lack of tools, which let users browse the original raw data and its raw signals in a smooth, interactive way as known from desktop applications. This can be of particular importance especially for *de novo* identification of molecules. Thus metabolomics web platforms, including MeltDB 2.0, are mostly useful for experiments with large numbers of chromatograms, which ask for the statistical comparison of sample groups.

Further developments of MeltDB should strive to include support for multi-stage (MS^n) data, another inevitable tool for proper *de novo* identification.

From a more global and systemic point of view into the future, the potential of integrated analysis with other omics data needs to be explored more intensively. As an example, the MeltDB 2.0 API allows the ProMeTra (Neuweger, 2009) software to map relative metabolite abundances to pathway maps, together with either proteome or transcriptome data.

In total, MeltDB has undergone substantial improvements in its capacity as a 'one-stop-shop' providing a wide spectrum of necessary tools to answer biological and statistical questions, beginning from GC-MS *raw data* files. The addition of supervised machine-learning tools now allows to directly apply gathered knowledge for classification purposes. Embedded in its powerful permission management system, MeltDB 2.0 delivers a comprehensive bioinformatics package for detailed, systemic metabolomics research projects.

8 STUDY: Multivariate GC-MS wheat data analysis

The increasing awareness of the benefits of healthy eating has tremendously risen the popularity of organic food - a development that was not least stirred up by the manifold food scandals grabbing the headlines in recent years. Directly resulting from this popularity but in particular from organic food's great market potential, there emerged a significant interest in the authenticity of food declared as organic (Capuano *et al.*, 2013). Metabolomics technologies have proven successful for several tasks of food authentication (Cubero-Leon *et al.*, 2014). This study aims to investigate the potential of metabolomics profiling techniques, bioinformatics and machine learning to distinguish organically grown wheat from conventionally grown wheat. To this end, a total of more than 300 gas chromatography-mass spectrometry (GC-MS) measurements from both types of treatments were recorded and analyzed. Samples comprised eleven different wheat cultivars from up to three different years, obtained from the DOK field trial in Switzerland (Mäder *et al.*, 2002). This comprehensive field trial compared organic and conventional farming systems, using strictly controlled conditions. In previous work Bonte *et al.* (2014) already presented metabolite profiling data obtained from the DOK wheat samples of the harvest year 2007. Röhlig and Engel (2010) have applied principal component analysis (PCA) and analysis of variance (ANOVA) to a very similar dataset. In the scope of this work, the DOK data basis from 2007 has been extended substantially by additionally analyzing samples from the 2009 and 2010 harvest years. The particular focus of this work was placed on the potential of machine learning methods as tools for automated data classification. Furthermore the new approach is metabolite-agnostic: It does not rely on correct metabolite identification and it does not rely on single biomarkers with significant level differences. The latter is a core advantage of this approach, as literature reveals that only slight (not significant) metabolite level changes can be accounted on the farming systems (Röhlig and Engel, 2010; Laursen *et al.*, 2011; Bonte *et al.*, 2014).

All GC-MS measurements were automatically preprocessed and then carefully annotated in the MeltDB 2.0 metabolomics analysis platform (Neuweger *et al.*, 2008; Kessler *et al.*, 2013). MeltDB allowed to apply a well-established routine

in high-dimensional molecular data analysis. After pre-processig (peak picking, normalization, profiling etc.) the data is represented as a table of dimension $n \times D$, with n = number of samples and D = signal dimension (i.e. the metabolic profile). The first aim is to search for hidden regularities, relationships, and correlations in the data. To this end, unsupervised learning, i.e. dimensional reduction can be applied. Concretely, the two unsupervised methods principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE, Van Der Maaten and Hinton (2008)) were used to investigate the inter- and intra-class variances in the entire dataset as well as in particular subsets of the data.

Second, the data was analyzed towards the question, if it can be classified into distinct semantic categories (like conventional / organic treatment in this case). Hence the two supervised machine learning methods random forests (RF, Breiman (2001)) and support vector machines (SVM, Vapnik (1999)) were applied. The overall aim was to establish a classifier to distinguish between organic and conventional wheat, despite the influences of the years of growth and different cultivars.

In the following, the analytical approach is described in detail, as well as the separation results for the investigated factors treatment, cultivar and year. All presented computational methods were implemented within the MeltDB 2.0 platform and can be applied to other datasets as well.

Table 8.1: Number of samples for each combination of factors 'farming system', 'year', and 'cultivar'.

		Cultivar											
Farming system	Year	Antonius	Caphorn	CCP	DJ 9714	Mont Calme 245	Probus	Rouge de Bordeaux	Runal	Sandomir	Scaro	Titlis	Σ
conventional	2007	7	8	7	7	8	7	7	8	6	6	8	160
organic		8	7	7	7	7	8	8	7	8	7	7	
conventional	2009								8				16
organic									8				
conventional	2010	8	8	7			8	8	7	8	7	8	137
organic		8	8	8			8	8	7	7	7	7	

8.1 Wheat sample preparation and GC-MS analysis

Wheat grains of up to 11 different cultivars originated from the DOK (D: biodynamic, O: bio-organic, K: 'konventionell' German for conventional, i.e. integrated, farming system) field trial, which is located at Therwil (7°33' E, 47°30' N) close to Basel (Switzerland). Detailed Information on the DOK long-term field trial is given by Mäder *et al.* (2002). Wheat grains of the cultivar Runal were analyzed from the three harvest years 2007, 2009 and 2010. In 2008 wheat was not grown in the trial. Further, the 10 wheat cultivars 'Rouge de Bordeaux', 'Mont Calme 245', 'Probus', 'CCP' (composite cross-population; for ease of reading CCP is referred to as a cultivar), 'Scaro', 'Sandomir', 'DJ 9714', 'Antonius', 'Caphorn' and 'Titlis' were integrated into the wheat plots of the long term trial of the harvest year 2007. In the 2010 cultivation period, cultivars 'Mont Calme 245' and 'DJ9714' were not available, leaving the remaining 8 cultivars mentioned previously for analysis in this work. A detailed description of the layout and design of the experiment comprising all winter wheat cultivars was published by Hildermann *et al.* in 2009.

Thus only some essential information about the DOK field trial is considered here. The trial comprises several organic and conventional farming systems, each system being repeated in four field plots. The experimental design was a split plot with systems as the main factor and wheat cultivars as the secondary factor. This study focuses on the analysis of the two farming systems biodynamic 2 (D), (henceforth, organic) and conventional (M). These two farming systems were quite different with respect to fertilization and further plant treatment (see below), but at the same time were still within the range of standard organic and conventional farming.

The organic system received composted manure and slurry at a fertilization level of 1.4 livestock units per hectare, equivalent to 66 kg N(total) ha^{-1}. Fertilization in the conventional system was done exclusively with mineral fertilizer at 140 kg N(total) ha^{-1}. Both farming systems also differed in plant protection practice. The conventional system followed the guidelines of integrated farming, using fungicides, insecticides and herbicides only if needed. The biodynamic farming allowed only mechanical plant treatments and indirect methods to control weeds, pests and diseases. Grains of both farming systems were harvested when completely ripe, with moisture content below 140 g kg^{-1}. Of each of the four individual field plots per agricultural system, one sample was taken for each cultivar and farming system. Before further experimental usage, grain material was stored at a constant temperature of 18 °C.

Cleaning of wheat samples from impurities and broken grains, grain storage, grinding and extraction as well as measurement of metabolites using GC-MS analysis was exactly performed as described by Bonte *et al.* (2014).

8.2 Data preprocessing and feature annotation

All data gained by GC-MS analysis were preprocessed and annotated within the MeltDB 2.0 metabolomics software platform. Peaks where obtained using the Warped Peak Detection tool. Retention indices were obtained semi-automatically, using MeltDB's RISimple tool and a manually defined list of expected retention times for each batch of measurements. Next, a profiling was run to annotate peaks that are common throughout multiple chromatograms, i.e. they have a similar retention index and a similar EIC spectrum. In a like manner all chromatograms where matched against reference spectra to annotate peaks as identified compounds where possible. However the subsequent approach does not rely on the identification of compounds, but rather uses it to limit the feature space to molecules of potential biological interest. The parameterizations for these processing tools can be found in Tab. 8.2.

Results from automated metabolite identification were revised manually to discard erroneous annotations, but also to manually create annotations that were missed in a few chromatograms only. Peaks that were missed in a minority of samples were requantified using the Warped Peak Detection tool. Subsequently all data in the obtained feature table was centered and scaled using R.

In total 313 samples were analyzed. From the years 2007, 2009, and 2010 these comprise 160, 16, and 137 samples, respectively. How many cultivars were available in each year is mentioned in section 8.1. For each combination of year and cultivar 13 to 16 samples, comprising duplicate metabolite extractions for grains from most field plots, were analyzed. One half of these was treated organically, and the other half was treated conventionally. A detailed listing of all samples is given in Tab. 8.1.

8.2.1 Unsupervised learning / dimensional reduction

The result of the pre-processing step is an $n \times D$ data table, where n=313 is the total number of samples (years, cultivars and treatments combined) and D=36 is the number of compounds consistently annotated in all these samples. Additionally, subsets of this data table, for example all samples from only one year or cultivar, have been analyzed too. Although it may be tempting to instantly apply supervised learning to the problem of classifying the data rows into conventional and biological treatment, it is advisable to first apply some information visualization in advance to avoid unpleasant black box effects and to gain a mental model of the data. Information visualization uses different data displays which are inspected by human experts to understand the data or to build hypotheses for the hidden structures in the data. These are the foundation for any subsequent attempt to apply supervised learning. Therefore displays obtained with two different dimen-

Table 8.2: Parameters that were applied for the preprocessing tools in MeltDB.

Tool	Description	Parameter	Value
Warped Peak Detection	Mexican-wavelet based peak detection, which can be rerun locally (at certain RT).	FWHM SN	7 10
RISimple	Detects and tags retention indices based on their characteristic spectra.	Ion filter	57, 71, 85, 99
Multiple Profiling	Gives peaks across chromatograms a common TAG if they are similar.	RT window	20-35 s
Reference List	Annotates peaks that match reference spectra, uses dot-product.	RT window	20 s

sional reduction techniques were inspected. PCA was applied since this is a well established statistics tool in high dimensional data analytics and is fully sufficient to understand data with a linear sub-structure. Since data stemming from systems biology experiments can not be expected to have such intrinsic linear structure, a supplementary method was used, which has been proposed in the field of machine learning: the t-SNE. In several real world applications for computational biology (Abdelmoula *et al.*, 2014; Bushati *et al.*, 2011; Jamieson *et al.*, 2010) t-SNE has shown to be capable of projecting non-linear data structure while well preserving the local features (i.e. neighborhoods) of the data.

The dimensional reductions were performed using the R statistical software (R_Development_Core_Team, 2011) and the `tsne` package by Donaldson (2012).

8.2.2 Supervised learning / classification

The same $n \times D$ data table was used to explore whether a machine learning algorithm such as the support vector machine (SVM, Vapnik (1999)) with a polynomial kernel (Karatzoglou *et al.*, 2004) can be trained to classify the data rows into

conventional and biological treatment. In a first step for each subset (e.g. data from one year only), the machine learning algorithm was trained and tested on 80 percent of the data (randomly selected). Afterwards, the remaining 20 percent out-of-the-bag data was used for validation, i.e. to finally evaluate the performance of the classifier constructed using the 80 percent of the data. To train and optimize the SVM, a parameter tuning was performed using a 25-fold resampling for Leave-Group-Out-Cross-Validation (LGOCV) of the training partition. For this cross validation, again 75 percent of the training partition were used for training, and 25 percent were used for validation in each iteration. The best set of parameters, which led to the best accuracy according to the LGOCV, was then once more validated on the 20 percent of the data that was kept back initially. The classification results on these latter 20 percent were evaluated in a confusion matrix to infer the accuracy of the trained SVM.

Using the very same subsets and partitions random forest (RF, Liaw and Wiener (2002)) was performed as well. The RF training was done with a 20-fold resampling and a parameter tune length of twelve.

Supervised machine learning methods were performed in R as well, using the `caret` package (Kuhn, 2008; Kuhn *et al.*, 2008).

8.3 Results

The strategy to first apply unsupervised and later supervised learning methods to the data gave insight into the structure of the investigated data: The dimensional reductions via PCA and t-SNE exposed how strongly the three different factors affect the clustering of the samples. This already guided the design and assessment of classifiers trained later on. The following sections present the results for both unsupervised and supervised learning methods.

8.3.1 Unsupervised learning / dimensional reduction

Both PCA (in the first two principal components, see Fig. 8.1) and t-SNE are capable of separating the presented wheat samples into clusters according to the factor year. Within one year the PCA will group samples according to cultivars, though with considerable overlap as can be seen in Fig. 8.2. Conversely, within one cultivar samples will be grouped according to the year (see Fig. 8.3). When one year and one cultivar are investigated in any combination, all data typically clusters into the two groups representing either dynamically or conventionally grown wheat. Most of these clusters show at least some overlap though. The t-SNE method is less applicable to smaller datasets and thus was applied to the complete data table only. Fig. 8.5 shows how t-SNE groups all samples by year

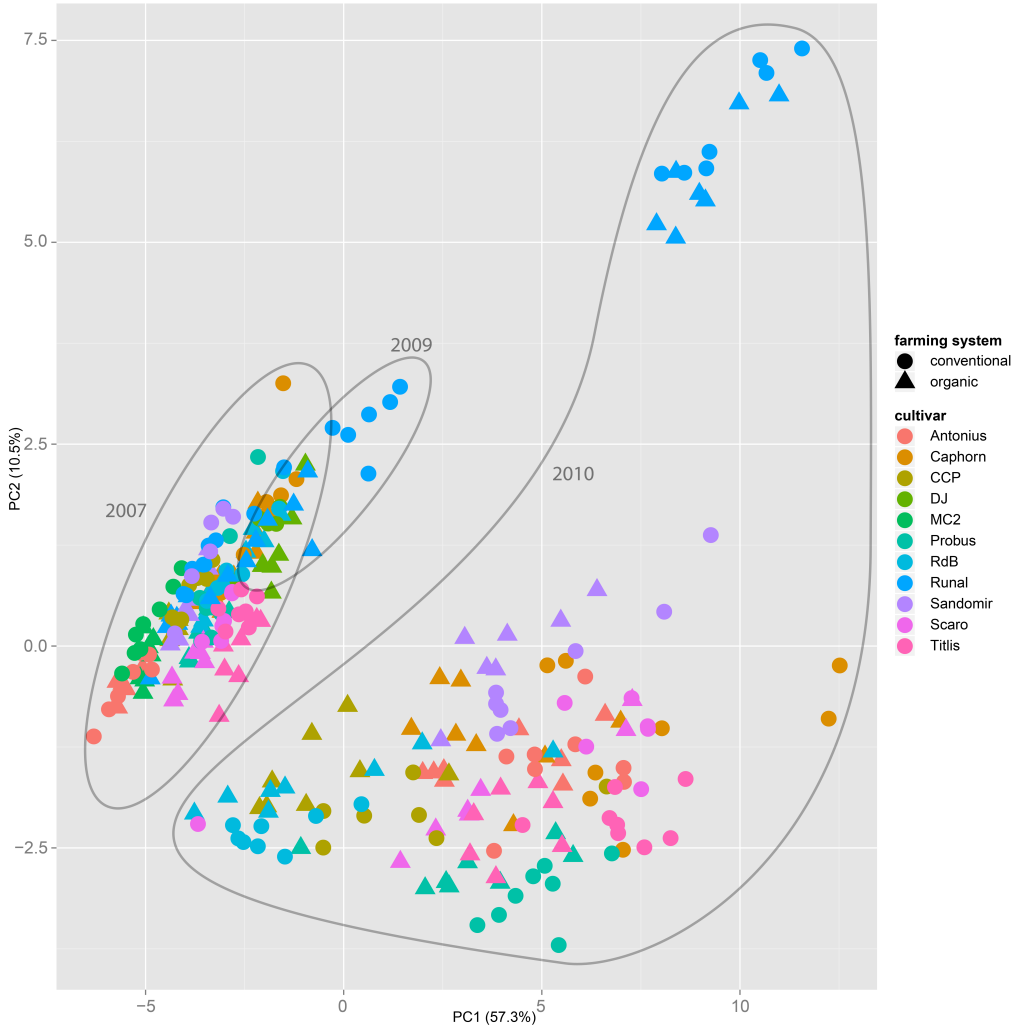

Figure 8.1: The principal component analysis on the entire dataset of all samples throughout all years, cultivars and treatments shows that the first two components mainly separate samples by the factor year. A separation by the factor farming system is not possible.

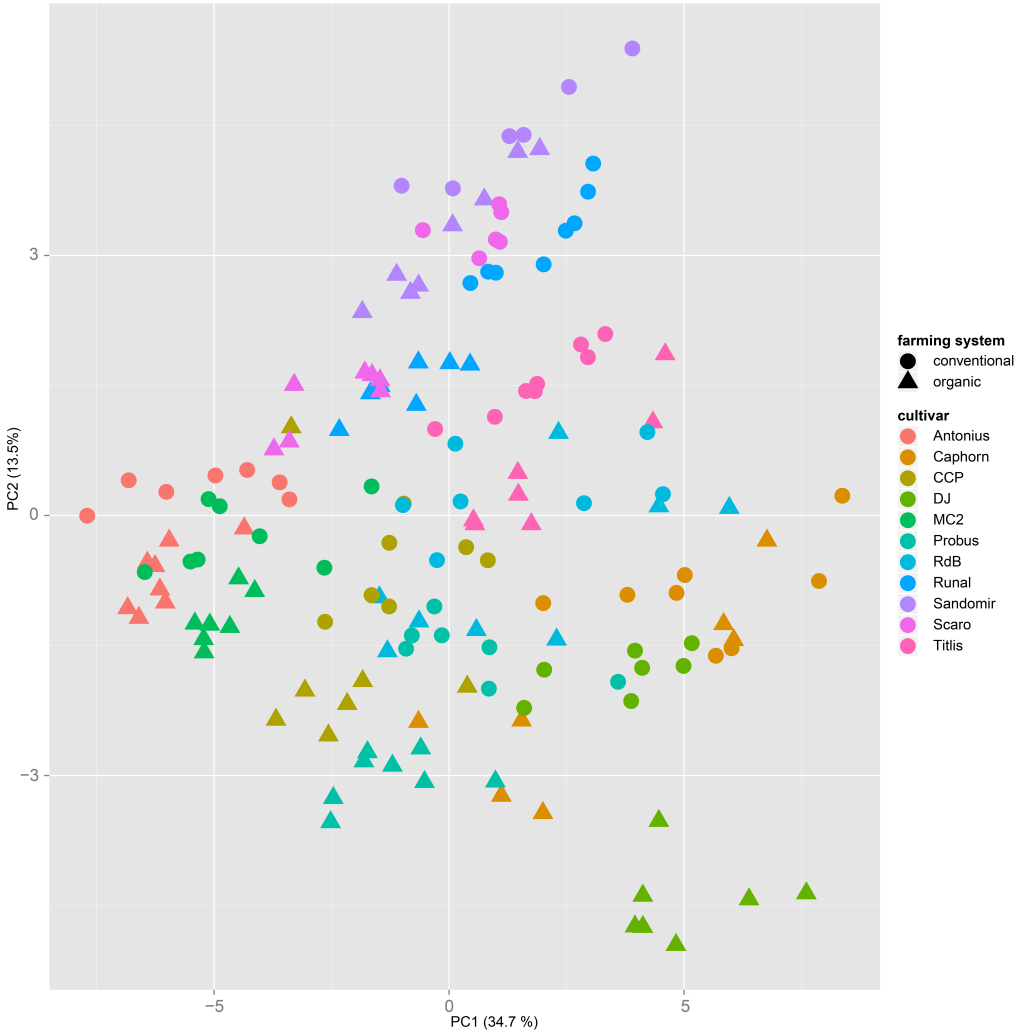

Figure 8.2: A principal component analysis performed on a dataset from one year only will mainly cluster samples by their cultivar, regardless of the applied farming system. This PCA is based on samples from the year 2007.

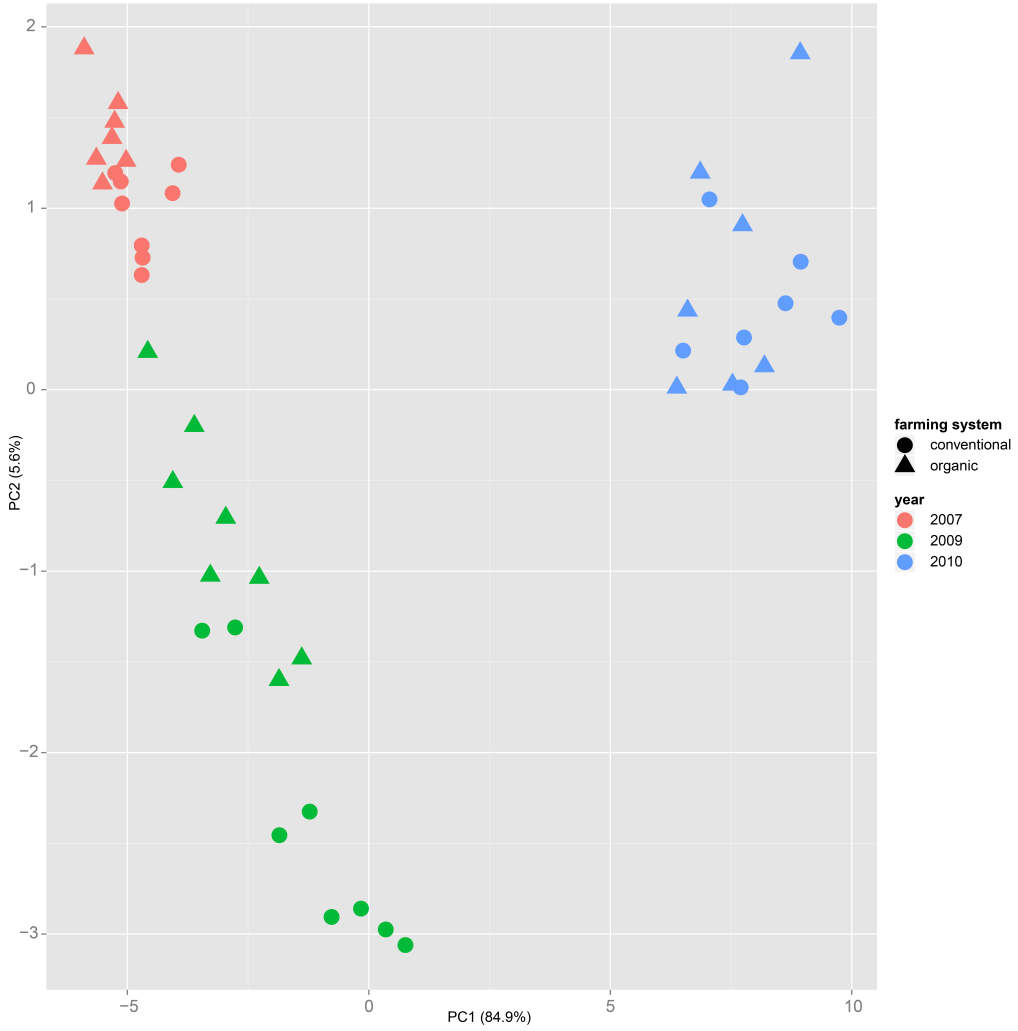

Figure 8.3: Similar to Fig. 8.1, in the principal component analysis on a dataset of only one cultivar - here 'Runal' is shown - the first principal components separate samples by factor year.

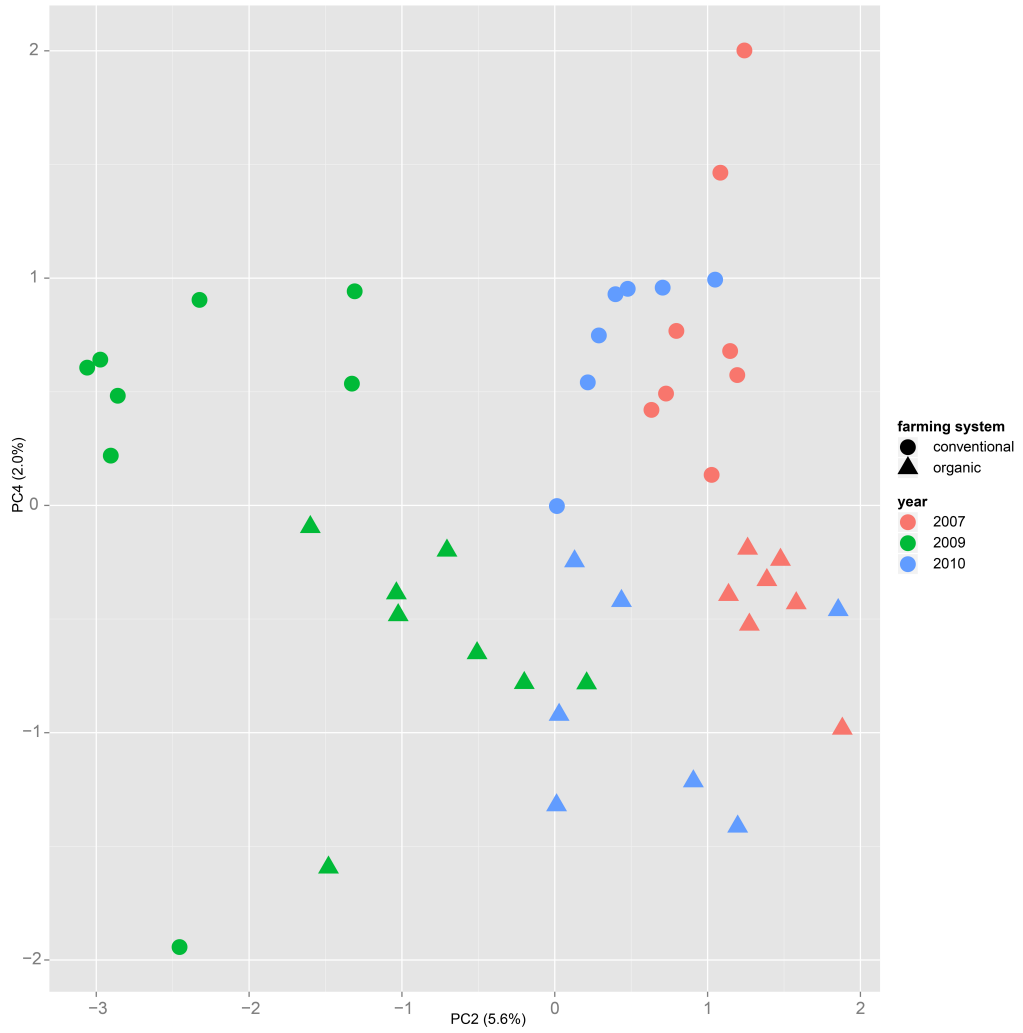

Figure 8.4: Plotting samples from one cultivar (here 'Runal') along the principal components two and four shows that a separation by farming system might be possible even though the main variance is caused by the factor year.

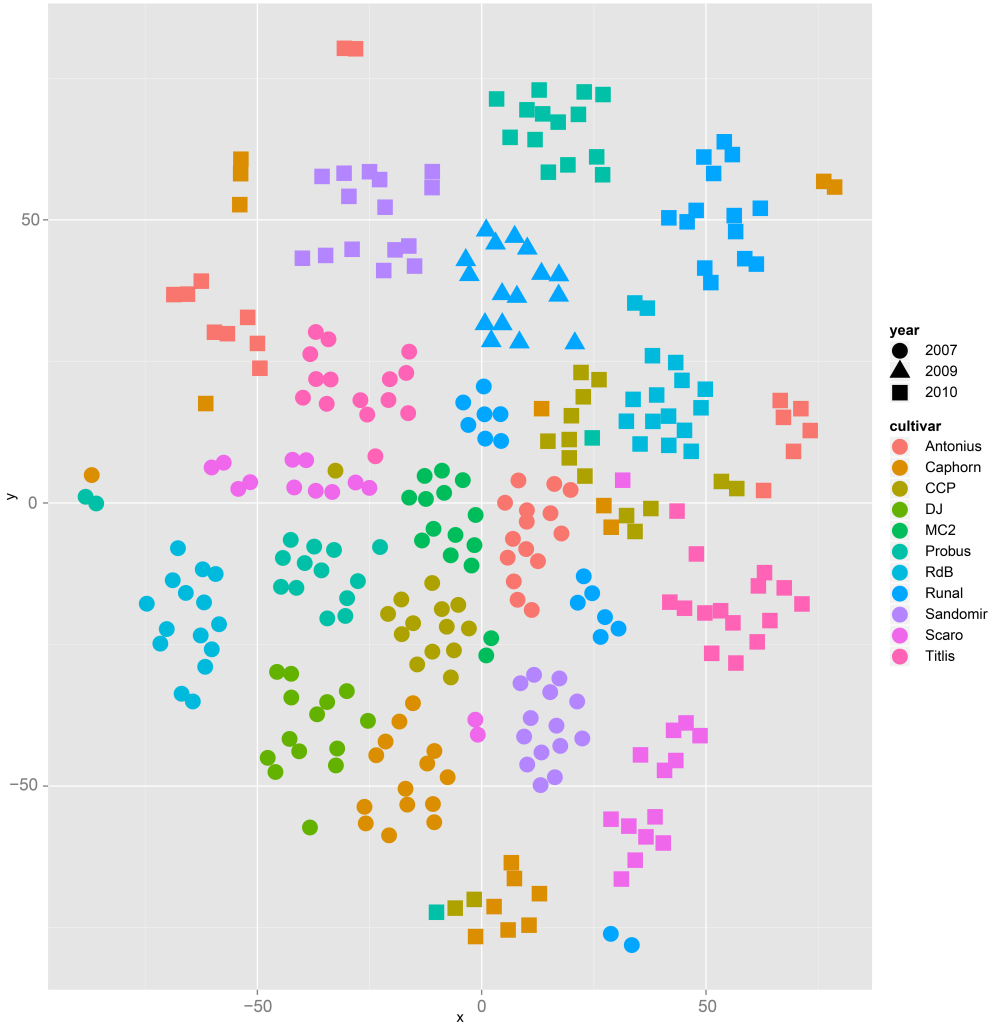

Figure 8.5: The t-SNE method applied to all samples results in clusters and sub clusters formed according to the factor year and cultivar, respectively.

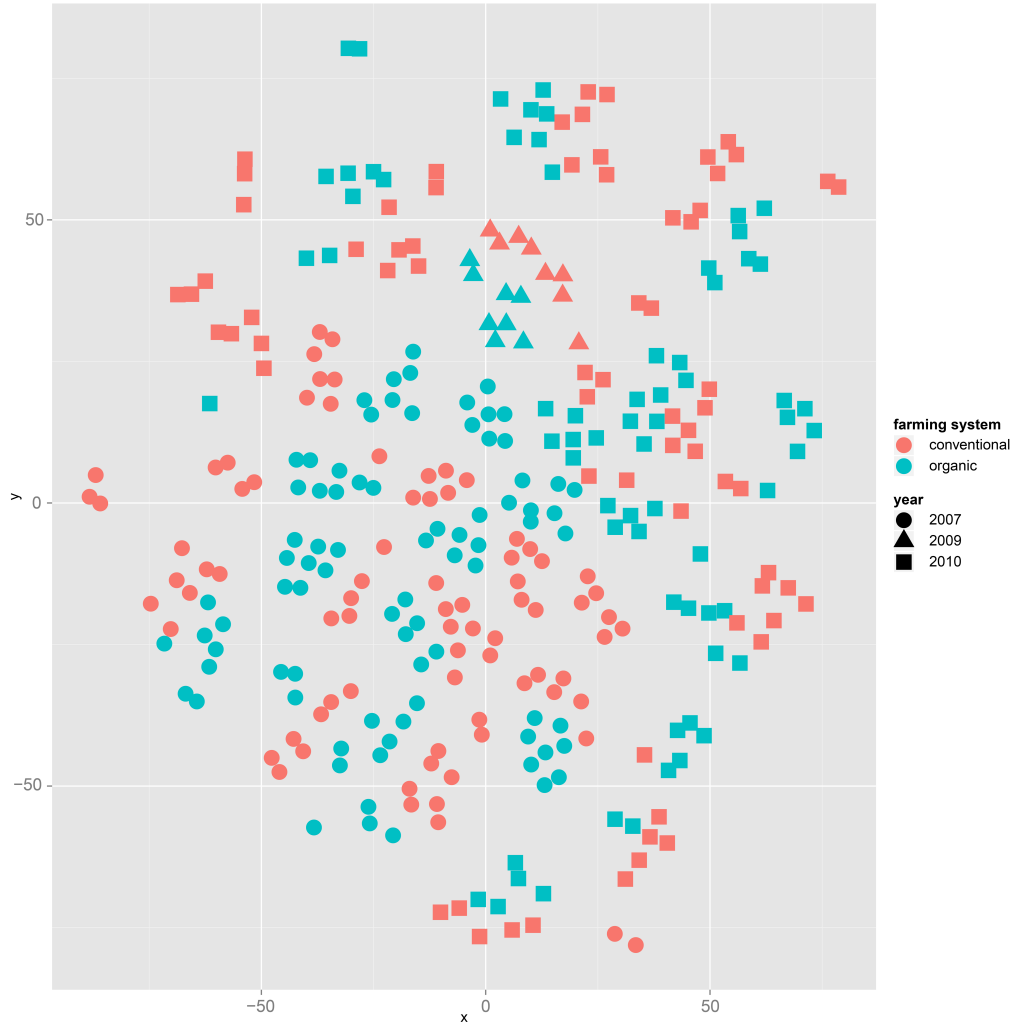

Figure 8.6: The same t-SNE result as in Fig. 8.5, but colored by farming system: Clusters representing cultivars form subclusters according to the factor farming system.

at first, and then into subclusters according to their cultivars. The latter subclusters themselves are again split in two groups each, which correspond to the two farming systems, as can be seen in Fig. 8.6. The Figs. 8.5 and 8.6 again visualize strikingly how the metabolic profile is mainly influenced by year, then by cultivar, and at least by the farming systems.

Nevertheless it is still obvious that the treatment caused measurable differences in the metabolic composition of the wheat samples. The first two principal components of the PCA in Fig. 8.3 already reveal that the clusters for the years 2007, 2009 and 2010 are split in themselves to form subclusters of conventionally and organically grown wheat. This points out that later principal components with different loadings may expose structures in the data that are mainly based on the factor treatment. Fig. 8.4 plots the second and fourth principal components of the PCA that has already been introduced in Fig. 8.3. In this case it is clearly visualized how the fourth principal component can be used to separate the samples according to the levels organic and conventional.

8.3.2 Supervised learning / classification

The results from the PCAs revealed that there are structures in the data that allow for a separation of conventionally and organically grown wheat. Even though the main clustering is driven by factor year, these clusters still form subclusters according to cultivar, which again are clustered by the two farming systems. These substructures suggest that SVMs can be constructed to win classifiers for the problem. In fact, SVMs trained and tested on the entire dataset (all years, all cultivars, both treatments) to classify by treatment reached an accuracy of 0.9032 (p-value = 1.486e-11, see Tab. 8.3) on the validation set. Even better accuracies can be observed when investigating subsets of the data (for example accuracy = 0.9677, p-value = 3.746e-08 within year 2007). But the smaller the subsets, the smaller the testing partitions, the less representative are any outcomes. Thus we will not trust the classifiers for in-cultivar or even in-year-and-cultivar problems to be flawless, even though in these cases accuracy values may approximate one easily.

The interesting question would be, if it is possible to obtain such a trained classifier from a number of (past) years, which can then be applied to classify samples from another (e.g. the present) year. This however turns out to be not possible on the basis of the available data from the three growing seasons. For example, when a SVM, trained on data from 2007, is applied to classify data from 2010 it performs with an accuracy of 0.5547, which is hardly favorable to plain guesses. The reason for this poor performance seems to be the massive influence of the seasonal conditions, i.e. the factor 'year'. This calls for continuing research using more samples from more years and cultivars to cover the molecular variance more appropriately. Estimations on the variable importance (Kuhn *et al.*, 2008) for the three years were

Table 8.3: Results of the Support Vector Machines, trained and tested on different subsets of all samples. Measures are given for the evaluation results and are based on the confusion matrix for classification as biological or conventional farming system.

Train	↦	Test	n_{Test}	Acc.	NIR[1]	p-value[2]	Sens.	Spec.	PPV[3]	NPV[4]
2007	↦	2007	31	0.9677	0.52	3.75E-08	1	0.9375	0.9375	1
2010	↦	2010	26	0.8846	0.5	4.40E-05	0.9231	0.8462	0.8571	0.9167
2007	↦	2010	137	0.5547	0.5	0.1333	0.2754	0.8382	0.6333	0.5327
2010	↦	2007	160	0.5562	0.51	0.1177	0.8101	0.3086	0.5333	0.625
2007 2009 2010	↦	2007 2009 2010	62	0.9032	0.5	1.49E-11	0.9032	0.9032	0.9032	0.9032

[1]No information rate: the larger class percentage; [2]Exact binomial test [Accuracy $>$ NIR]; [3]Positive predictive value; [4]Negative predictive value

calculated based on the SVM results and published in Kessler *et al.* (2015). Here it is striking that e.g. *myo*-inositol, which has previously been reported as a potential marker for farming systems (Röhlig and Engel, 2010; Bonte *et al.*, 2014), was most important for classification in 2007 but almost least important in 2009 and 2010. Such inhomogeneous variable importances additionally suggest a year-by-year strategy for training and classification.

Table 8.3 summarizes the SVM results. Please note that classification results for year 2009 are not reported here: with only one cultivar (Runal) and thus only 16 samples the subset is too small to generate reliable results. The 2009 samples are part of the analysis of the entire dataset, though.

Overall random forest (RF) as described in the methods section led to similar classification results, but showed slightly lower accuracies in the in-year subsets. Thus no detailed results are shown. However, random forest should not be ignored as a potential alternative for support vector machines in this scope.

8.4 Discussion

The main goal of this study was to investigate whether a classification of organically and conventionally grown wheat can be done, based on GC-MS metabolite measurements of wheat grains from different years and cultivars. Results from the unsupervised machine learning methods PCA and t-SNE show that the strongest variation in the data can be found in samples from different years. This may be in part due to different environmental influences and also due to systematic errors

that inevitably will occur in analyses from different years. Other studies report on the same obstructive effects (Röhlig and Engel, 2010; Laursen *et al.*, 2011). On the other hand though, this allows to extend the data basis every year. This demands robust classifiers that are able to cope with these kinds of problems, besides 'distracting' factors like year and cultivar. Further studies will additionally have to consider geographical influences on the metabolic composition of wheat grains.

Peaks from all 313 samples have been carefully annotated to achieve 36 consistently quantified features throughout the entire data set. These have first been explored with dimensional reduction methods like PCA and t-SNE to find the predominant structures in the data table. Then, supervised machine learning methods have been trained and applied to investigate in how far classifiers for organically and conventionally grown wheat can be created.

The considerably strong differences in samples from different years make it impossible though, to apply a classifier that was trained using data from year a_1 to distinguish data from another year a_2. To create a classifier for any year a_x, data from this a_x must be part of the training data set. PC analyses also suggest that it will be beneficial to concentrate on one cultivar or to have a broad data basis of many cultivars to cover variances that derive from this factor.

Support vector machines trained and applied on all samples from the same year, as well as SVMs trained and tested on all years, performed with high accuracies above or close to 0.9. This clearly outperforms the ability of PCA to separate samples according to the applied farming system, unless samples derive from the same cultivar. For comparison, random forests (RF) (Breiman, 2001) have been applied as an alternative for SVMs for classification. Random forests have the advantage to be much faster and more efficient than SVMs and they have the potential to offer some insight into the semantics of the decision function, but the parameters are more difficult to optimize. However, the classification performances were only slightly different from those obtained with SVMs and inferior in in-year analyses.

The here presented machine learning tools are not meant to substitute traditional statistical methods, such as ANOVA, but provide a metabolite-agnostic approach for sample classification where reliable biomarkers are not known. Additionally, they may contribute a starting point for focused statistical analyses of single compounds that appear promising according to the computed variable importance estimations.

An analytical approach that aims more for specific compounds as biological markers can be found in the publication of Bonte *et al.* (2014), where more traditional statistical methods have been applied. The methods presented in the manuscript at hand do not depend on the identification of compounds or the determination of the biological meaning of any features. The approach rather relies on a consistently annotated data set. Nevertheless it is constructive to do compound identification to be able to base further biomarker research on these studies.

Additionally, reducing the feature set to differentiating but also verified biological compounds minimizes the risk of systematic errors through background noise. The integration of the discussed approaches might finally lead to a set of metabolites that can be used as reliable biomarkers for conventional or biodynamic farming systems.

9 Discussion and Conclusion

The correct identification of metabolites is still the major bottleneck in metabolomics and ion deconvolution algorithms are one important tool to overcome this limitation. However, ion deconvolution heavily relies on expert knowledge and cannot be done automatically *per se*. There are too many ambiguities which cannot be resolved reliably by any software yet. The ALLocator online platform was the first software to provide in-depth interactive data exploration and curation (IDEC) in the domain of LC-ESI-MS ion deconvolution. Not only it brings the ALLocatorSD algorithm for the automated allocation and association of ion peaks into overlapping pseudo spectra, but it also provides interactive tables and visualizations for the IDEC of these spectra. The overall semi-automated workflow gives full control to the user and provides tools to confirm, deny, or alter pseudo spectra, which allows to profit from both: the algorithm's automation and the user's expert knowledge.

The complete, here presented semi-automated preprocessing workflow is available for both algorithms, ALLocatorSD (Kessler *et al.*, 2014) and CAMERA (Kuhl *et al.*, 2012), and it comprises feature finding, deisotoping, and pseudo spectra allocation. This boils the complex raw data sets down to the entities that finally matter for interpretation and statistics: Metabolites.

The in-source fragmentation is not only causing *noise*, which makes ion deconvolution necessary in the first place, but it can also be used as a source of information. When properly exploited, neutral losses help to identify the correct molecular formula by filtering out false positives which cannot explain the respective loss.

The same is true for U-^{13}C SIL experiments. Ion mass differences between U-^{12}C and U-^{13}C isotopologues can directly be used to defer the number of carbon atoms in the ions. In the small molecules domain this often narrows down the number of candidate formulas to one, when combined with fragment information and the 'Seven Golden Rules' by Kind and Fiehn (2007).

A possible future extension to the ALLocatorSD algorithm as well as to the filtering of molecular formulas is the matching of isotopic patterns. Each molecular composition results in certain intensity ratios for the monoisotopic and all following isotopic peaks, which can be calculated from the natural abundances of the isotopes of contributing elements. The different ions (adducts and fragments) of the same metabolite should thus have similar isotopic patterns. This could be used to assess the probability of two coeluting ions to derive from the same metabolite.

Furthermore, for any candidate molecular formula for a putative metabolite the theoretical isotopic pattern can be calculated and then could be matched against the measured isotopic pattern. This could be used to filter out molecular formulas that do not fit the measured isotopic patterns. The use of isotopic pattern fits heavily relies on two prerequisites though: First, a measured isotopic pattern has to be available, which is only given for the more abundant ions where even the isotopic peaks exceed the detection limit. Second, the intensities of all peaks in the isotope pattern have to have a good accuracy, which varies between different types of instruments and instruments of different vendors.

While the advantages of the semi-automated processing in ALLocator have been presented extensively in this thesis, the obvious drawback of any semi-automation is the necessary manual work. ALLocator thus allows to create custom reference lists of MS pseudo spectra, preserving the once made effort and invested time to curate the different ions deriving from specific compounds. These reference lists can be applied to other measurements in order to dereplicate - i.e. to reproduce the same allocation of ions into pseudo spectra and to annotate the known compounds.

The application study nicely comprises the power of the ALLocatorSD algorithm. Taking (γ-)glutamyl-valine as an example, where the deconvolution not only supported data reduction by allocating all ions, isotopologues and isotopic peaks, but also by filtering out false positive molecular formulas and even by allowing for the distinction of the $\alpha-$ and $\gamma-$ isomers, based on the identified neutral loss of NH_3. Especially the latter represents a depth of insight that is typically considered to require MS/MS analysis.

In total, the ALLocator approach to LC-ESI-MS data, exploiting all the information available to improve confidence in annotations, helps to close the gap between MS^1 and MS^2 analyses.

The development of MeltDB 2.0 advanced the handling of 100s of analyses in a single experiment. That mainly required improvement in two aspects: (a) the simplification of the annotation of common features across many analyses from multiple batches; and (b) interactive visualizations and statistics with on-the-fly filtering methods for their exploration.

The caveat that comes with the processing of analyses from multiple batches is the introduction of higher complexity in retention time shifts, matrix effects, and variations in the overall intensities. The benefit of combining many analyses from multiple batches into a single workflow for preprocessing and annotation is the prospect of well-founded statistical statements and the chance to train more robust classifiers for more complex classification problems.

To this end, the consistent annotation of features across large sets of measurements and interactive visualizations for statistical analyses have been implemented.

Furthermore, the MeltDB 2.0 software was extended with functionality for the training and application of machine-learning classifiers.

In the application study on GC-MS wheat analyses, investigating the influences of cultivation types, wheat cultivars, and the years of cultivation, these newly integrated machine learning tools were applied. More than 300 analyses, derived from samples grown and measured in three different years, were subjected to reference list matching in order to achieve a high quality dataset including 36 consistently annotated metabolic features. The study revealed the structure of the dataset, i.e. how clusters are formed according to the differently influential factors. It also showed the potential of the workflow, combining unsupervised and supervised learning methods, to deal with complex classification problems.

Both ALLocator and MeltDB are freely accessible through the Generalized Project Management System (GPMS) developed by the Bioinformatics Resource Facility at CeBiTec, Bielefeld University. Building on the GPMS and the nature of online platforms, researchers have the option to easily share their experiments with other users. This enables team members, ideally having different backgrounds, to conveniently work in distributed collaborations.

ALLocator and MeltDB constitute the two substantial pillars of an extensive, interactive and collaborative online software platform for the computational analysis of metabolomics data acquired with chromatography-hyphenated mass spectrometry.

10 Contributions to computational metabolomics

This chapter is intended to provide a concise overview of the most important contributions to computational metabolomics that are presented in this thesis. Each of these tools are publicly accessible through one of the web platforms ALLocator or MeltDB 2.0

ALLocatorSD: U-^{13}C-SIL capable spectra deconvolution

In addition to comprising adducts and neutral losses of measured compounds to pseudo spectra the ALLocator SD algorithm identifies isotope clusters that derive from U-^{13}C labeled moieties and associates them to their ^{12}C counterparts. That is not only one more step for data reduction, but the nominal mass difference from ^{12}C and ^{13}C monoisotopic peaks (i.e. the number of carbon atoms) is applied to further improve the identification of unexpected neutral losses and to support mass decomposition (see section on spectrum-driven mass decomposition below).

More details in sections: 2.3.6 Ionization methods, 2.5.1 Spectra deconvolution, 2.6.5 Stable isotope labeling, 5.3 Spectra deconvolution algorithm, 6 STUDY: *Amino-acid profiling in C. glutamicum strains.*

ALLocator web platform: Manual curation of ambiguous fragmentation spectra

The convolution of LC-ESI-MS is a problem that cannot be solved unambiguously by any tool yet. E.g. it is often not clear, whether a certain peak represents a neutral loss of molecule A, an adduct of molecule B, or the pseudo-molecular ion of molecule C, when all these possibilities are backed by further peaks and each combination forms a valid pseudo spectrum. ALLocator introduces a new user interface which allows to interactively resolve conflicting pseudo spectra by 'claiming'

ion peaks for the correct spectrum, or removing them from an incorrect spectrum, or to add further peaks that correlate well over retention time.

More details in sections: 2.5.1 Spectra deconvolution, 5.3 Spectra deconvolution algorithm, 5.5 Data curation, 6 STUDY: *Amino-acid profiling in C. glutamicum strains.*

ALLocator web platform: Spectrum-driven mass decomposition

The struggle for better mass decomposition is typically thought to be limited by the mass spectrometers precision and it is only supported by (heuristic) chemical rules (Kind and Fiehn, 2007) to narrow down the list of putative candidates. The ALLocator web interface not only allows to activate and deactivate these rules for mass decomposition, but additionally includes the information from neutral losses and even the U-^{13}C-SIL. This often allows to elucidate the correct molecular formula unequivocally and increases the level of confidence according to metabolite identification reporting standards (Fiehn *et al.*, 2007; Schymanski *et al.*, 2014; Sumner *et al.*, 2014). In their combination, ALLocatorSD and the spectrum-driven mass decomposition leverage in-source fragmentations from undesirable obstructions to a powerful source of information.

More details in sections: 2.3.6 Ionization methods, 2.5.1 Spectra deconvolution, 2.5.3 Mass decomposition, 2.5.4 Metabolite identification, 2.6.5 Stable isotope labeling, 5.3 Spectra deconvolution algorithm, 5.4.4 Mass decomposition and search by molecular formula (ChemSpider), 6 STUDY: *Amino-acid profiling in C. glutamicum strains.*

MeltDB 2.0: Handling hundreds of chromatograms efficiently

The MeltDB web platform can deal with large experiments containing hundreds of chromatograms, and it offers a vast toolbox of statistics that can be applied. However, with increasing data volume both the navigation through experiments and the calculation of statistics and their static plots became cumbersome. New interactive visualizations for statistics and dynamic features in the main experiment view reduced the required data access (i.e. database queries) and sped up the general handling. This way it is now possible to efficiently investigate experiments that contain several hundreds of chromatograms.

More details in sections: 7.2 General workflow and integrated features, 7.2.3 User inter-

faces for all levels of data abstraction, 7.2.4 Statistics and data mining, 8 STUDY: *Multivariate GC-MS wheat data analysis.*

MeltDB 2.0: Integrated access to machine learning algorithms

The possibility to handle larger sets of chromatograms (i.e. samples) unlocked the doors to meaningful machine learning experiments. Now it is possible to generate training sets that are large enough to yield the capabilities of machine learning algorithms. A new interface was developed that makes these algorithms accessible directly from the MeltDB web platform, by selecting and grouping chromatograms to train classifiers that can be directly applied to classify unknowns.

More details in sections: 7.2 General workflow and integrated features, 7.2.3 User interfaces for all levels of data abstraction, 7.2.4 Statistics and data mining, 8 STUDY: *Multivariate GC-MS wheat data analysis.*

11 Bibliography

Abdelmoula, W. M., Škrášková, K., Balluff, B., Carreira, R. J., Tolner, E. a., Lelieveldt, B. P. F., van der Maaten, L., Morreau, H., van den Maagdenberg, A. M. J. M., Heeren, R. M. a., McDonnell, L. a., and Dijkstra, J. (2014). Automatic generic registration of mass spectrometry imaging data to histology using nonlinear stochastic embedding. *Analytical Chemistry*, **86**(18), 9204–9211.

Annesley, T. M. (2003). Ion suppression in mass spectrometry. *Clinical chemistry*, **49**(7), 1041–4.

Arita, M. (2004). Computational resources for metabolomics. *Briefings in functional genomics & proteomics*, **3**(1), 84–93.

Baran, R., Bowen, B. P., Bouskill, N. J., Brodie, E. L., Yannone, S. M., and Northen, T. R. (2010). Metabolite Identification in Synechococcus sp. PCC 7002 Using Untargeted Stable Isotope Assisted Metabolite Profiling. *Analytical chemistry*, **82**(21), 9034–9042.

Baran, R., Ivanova, N. N., Jose, N., Garcia-pichel, F., Kyrpides, N. C., Gugger, M., and Northen, T. R. (2013). Functional Genomics of Novel Secondary Metabolites from Diverse Cyanobacteria Using Untargeted Metabolomics. *Marine Drugs*, **11**(10), 3617–3631.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, **57**(1), 289–300.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S., and Sumner, L. W. (2004). Potential of metabolomics as a functional genomics tool. *Trends in plant science*, **9**(9), 418–25.

Bolten, C. J., Kiefer, P., Letisse, F., Portais, J.-C., and Wittmann, C. (2007). Sampling for metabolome analysis of microorganisms. *Analytical chemistry*, **79**(10), 3843–9.

Bonte, A., Neuweger, H., Goesmann, A., Thonar, C., Mäder, P., Langenkämper, G., and Niehaus, K. (2014). Metabolite profiling on wheat grain to enable a distinction of samples from organic and conventional farming systems. *Journal of the science of food and agriculture*, **94**(13), 2605–2612.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

Bristow, A. W. T., Webb, K. S., Lubben, A. T., and Halket, J. (2004). Reproducible product-ion tandem mass spectra on various liquid chromatography/mass spectrometry instruments for the development of spectral libraries. *Rapid communications in mass spectrometry*, **18**(13), 1447–54.

Bueschl, C., Kluger, B., Berthiller, F., Lirk, G., Winkler, S., Krska, R., and Schuhmacher, R. (2012). MetExtract: a new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research. *Bioinformatics (Oxford, England)*, **28**(5), 736–8.

Bueschl, C., Kluger, B., Lemmens, M., Adam, G., Wiesenberger, G., Maschietto, V., Marocco, A., Strauss, J., Bödi, S., Thallinger, G. G., Krska, R., and Schuhmacher, R. (2013a). A novel stable isotope labelling assisted workflow for improved untargeted LC-HRMS based metabolomics research. *Metabolomics*, **10**(4), 754–769.

Bueschl, C., Krska, R., Kluger, B., and Schuhmacher, R. (2013b). Isotopic labeling-assisted metabolomics using LC-MS. *Analytical and Bioanalytical Chemistry*, **405**(1), 27–33.

Bueschl, C., Kluger, B., Neumann, N. K., Doppler, M., Maschietto, V., Thallinger, G. G., Meng-Reiterer, J., Krska, R., and Schuhmacher, R. (2017). MetExtract II: A Software Suite for Stable Isotope-Assisted Untargeted Metabolomics. *Analytical Chemistry*, **89**(17), 9518–9526.

Bushati, N., Smith, J., Briscoe, J., and Watkins, C. (2011). An intuitive graphical visualization technique for the interrogation of transcriptome data. *Nucleic acids research*, **39**(17), 7380–9.

Capuano, E., Boerrigter-Eenling, R., van der Veer, G., and van Ruth, S. M. (2013). Analytical authentication of organic products: An overview of markers. *Journal of the Science of Food and Agriculture*, **93**(1), 12–28.

Carroll, A. J., Badger, M. R., and Harvey Millar, a. (2010). The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC bioinformatics*, **11**, 376.

Castillo, S., Gopalacharyulu, P., Yetukuri, L., and Orešič, M. (2011). Algorithms and tools for the preprocessing of LCâMS metabolomics data. *Chemometrics and Intelligent Laboratory Systems*, **108**(1), 23–32.

Choi, B. K., Hercules, D. M., and Gusev, a. I. (2001). Effect of liquid chromatography separation of complex matrices on liquid chromatography-tandem mass spectrometry signal suppression. *Journal of chromatography. A*, **907**(1-2), 337–42.

Chokkathukalam, A., Jankevics, A., Creek, D. J., Achcar, F., Barrett, M. P., and Breitling, R. (2013). mzMatch-ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. *Bioinformatics (Ox-*

ford, England), **29**(2), 281–3.

Creek, D. J., Dunn, W. B., Fiehn, O., Griffin, J. L., Hall, R. D., Lei, Z., Mistrik, R., Neumann, S., Schymanski, E. L., Sumner, L. W., Trengove, R., and Wolfender, J.-L. (2014). Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics*, **10**(3), 350–353.

Cubero-Leon, E., Peñalver, R., and Maquet, A. (2014). Review on metabolomics for food authentication. *Food Research International*, **60**, 95–107.

Danielsson, R., Bylund, D., and Markides, K. E. (2002). Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatographyâmass spectrometry. *Analytica chimica acta*, **454**(2), 167–184.

de Hoffmann, E. and Stroobant, V. (2001). *Mass spectrometry : principles and applications*. Wiley, Chichester New York, 2nd edition.

de Laeter, J. R., Böhlke, J. K., De Bièvre, P., Hidaka, H., Peiser, H. S., Rosman, K. J. R., and Taylor, P. D. P. (2003). Atomic weights of the elements. Review 2000 (IUPAC Technical Report). *Pure and Applied Chemistry*, **75**(6), 683–799.

Donaldson, J. (2012). tsne: T-distributed Stochastic Neighbor Embedding for R (t-SNE).

Ettre, L. (1994). New, unified nomenclature for chromatography. *Chromatographia*, **38**(7), 521–526.

Ettre, L. and Sakodynskii, K. (1993a). M . S . Tswett and the Discovery of Chromatography I: Early Work (1899-1903). *Chromatographia*, **35**(3/4), 223–231.

Ettre, L. and Sakodynskii, K. (1993b). M.S. Tswett and the Discovery of Chromatography II: Completion of the Development of Chromatography. *Chromatographia*, **35**(5/6), 329–338.

Fernie, A., Trethewey, R., and Krotzky, A. (2004). Metabolite profiling: from diagnostics to systems biology. *Nature Reviews Molecular Cell Biology*, **5**, 763–769.

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant molecular biology*, **48**(1-2), 155–71.

Fiehn, O., Robertson, D., Griffin, J., Werf, M., Nikolau, B., Morrison, N., Sumner, L. W., Goodacre, R., Hardy, N. W., Taylor, C., Fostel, J., Kristal, B., Kaddurah-Daouk, R., Mendes, P., Ommen, B., Lindon, J. C., and Sansone, S.-A. (2007). The metabolomics standards initiative (MSI). *Metabolomics*, **3**(3), 175–178.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-f., Dougherty, B. A., Merrick, J. M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-l., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T.,

Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., Mcdonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., and Ventert, J. C. (1995). Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *Science*, **269**(July), 496–512.

Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Research*, **13**(2), 244–253.

Freiser, H. and Nancollas, G. (1987). *Compendium of Analytical Nomenclature - The Orange Book*. Blackwell Science, Oxford, 2nd edition.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10), R80.

German, J. B., Hammock, B. D., and Watkins, S. M. (2005). Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, **1**(1), 3–9.

Giavalisco, P., Köhl, K., Hummel, J., Seiwert, B., and Willmitzer, L. (2009). 13C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Analytical chemistry*, **81**(15), 6546–51.

Gofieau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Louis, E., Mewes, H., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. (1996). Life with 6000 Genes. *Science*, **274**(5287), 546–67.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. *Molecular and cellular biology*, **19**(3), 1720–1730.

Halket, J. M., Przyborowska, A., Stein, S. E., Mallard, W. G., Down, S., and Chalmers, R. a. (1999). Deconvolution gas chromatography/mass spectrometry of urinary organic acids–potential for pattern recognition and automated identification of metabolic disorders. *Rapid Communications in Mass Spectrometry*, **13**(4), 279–84.

Hansen, M. A. E., Villas-Bôas, S. G., Roessner, U., Hansen, M. A. E., Smedsgaard, J., and Nielsen, J. (2006). Data Analysis. In D. M. Desiderio and N. M. M. Nibbering, editors, *Metabolome Analysis: An Introduction*, chapter 5, pages 146–187. John Wiley & Sons, Inc., Hoboken.

Harrison, A. G. (2003). Fragmentation reactions of protonated peptides containing

glutamine or glutamic acid. *Journal of Mass Spectrometry*, **38**(2), 174–87.

Hasegawa, M. and Matsubara, I. (1978). Gamma-Glutamylpeptide formative activity by Corynebacterium glutamicum by the reverse reaction of the gamma-glutamylpetide hydrolytic enzyme. *Agricultural and Biological Chemistry*, **42**(2).

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York Inc., New York, 2nd edition.

Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone, S. A., Griffin, J. L., and Steinbeck, C. (2013). MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, **41**(D1), 781–786.

Hegeman, A. D., Schulte, C. F., Cui, Q., Lewis, I. a., Huttlin, E. L., Eghbalnia, H., Harms, A. C., Ulrich, E. L., Markley, J. L., and Sussman, M. R. (2007). Stable isotope assisted assignment of elemental compositions for metabolomics. *Analytical chemistry*, **79**(18), 6912–21.

Hertz, H. S., Hites, R. A., and Biermann, K. (1971). Identification of Mass Spectra by Computer-Searching a File of Known Spectra. *Analytical chemistry*, **43**(6), 681–691.

Hildermann, I., Thommen, A., Dubois, D., Boller, T., Wiemken, A., and Mäder, P. (2009). Yield and baking quality of winter wheat cultivars in different farming systems of the DOK long-term trial. *Journal of the Science of Food and Agriculture*, **89**(14), 2477–2491.

Hoffmann, N. and Stoye, J. (2009). ChromA: signal-based retention time alignment for chromatography-mass spectrometry data. *Bioinformatics (Oxford, England)*, **25**(16), 2080–1.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, **6**(2), 65–70.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., and Nishioka, T. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, **45**(7), 703–714.

Huang, X., Chen, Y.-J., Cho, K., Nikolskiy, I., Crawford, P. a., and Patti, G. J. (2014). X13CMS: global tracking of isotopic labels in untargeted metabolomics. *Analytical chemistry*, **86**(3), 1632–9.

Ideker, T., Galitski, T., and Hood, L. (2001). A New Approach to Decoding Life: Systems Biology. *Annual Review of Genomics and Human Genetics*, **2**, 343–372.

Jaitly, N., Mayampurath, A., Littlefield, K., Adkins, J. N., Anderson, G. A., and Smith, R. D. (2009). Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC bioinformatics*, **10**, 87.

Jamieson, A. R., Giger, M. L., Drukker, K., Li, H., Yuan, Y., and Bhooshan, N. (2010). Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian eigenmaps and t-SNE. *Medical Physics*, **37**(1), 339–51.

Juan, Y.-F. (2003). *Statistics::TTest*. CPAN module version 1.1.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, **11**(9).

Kastenmüller, G., Römisch-Margl, W., Wägele, B., Altmaier, E., and Suhre, K. (2011). meta P- Server : A Web-Based Metabolomics Data Analysis Tool. *Journal of Biomedicine and Biotechnology*, **2011**, 1–7.

Katajamaa, M. and Oresic, M. (2005). Processing methods for differential analysis of LC/MS profile data. *BMC bioinformatics*, **6**, 179.

Katajamaa, M., Miettinen, J., and Oresic, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics (Oxford, England)*, **22**(5), 634–6.

Keller, B. O., Sui, J., Young, A. B., and Whittal, R. M. (2008). Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta*, **627**(1), 71–81.

Kessler, N., Neuweger, H., Bonte, A., Langenkämper, G., Niehaus, K., Nattkemper, T. W., and Goesmann, A. (2013). MeltDB 2.0-advances of the metabolomics software system. *Bioinformatics*, **29**(19), 2452–2459.

Kessler, N., Walter, F., Persicke, M., Albaum, S. P., Kalinowski, J., Goesmann, A., Niehaus, K., and Nattkemper, T. W. (2014). ALLocator: An Interactive Web Platform for the Analysis of Metabolomic LC-ESI-MS Datasets, Enabling Semi-Automated, User-Revised Compound Annotation and Mass Isotopomer Ratio Analysis. *PloS one*, **9**(11), e113909.

Kessler, N., Bonte, A., Albaum, S. P., Mäder, P., Messmer, M., Goesmann, A., Niehaus, K., Langenkämper, G., and Nattkemper, T. W. (2015). Learning to Classify Organic and Conventional Wheat - A Machine Learning Driven Approach Using the MeltDB 2.0 Metabolomics Analysis Platform. *Frontiers in Bioengineer-*

ing and Biotechnology, **3**(35), 1–10.

Kind, T. and Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics*, **7**, 234.

Kind, T. and Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, **8**, 105.

Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S., and Fiehn, O. (2009). FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical chemistry*, **81**(24), 10038–48.

Kopka, J., Schauer, N. N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., Dörmann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A. R., and Steinhauser, D. (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics (Oxford, England)*, **21**(8), 1635–1638.

Kuhl, C. and Tautenhahn, R. (2010). LC-MS Peak Annotation and Identification with CAMERA. pages 1–14.

Kuhl, C., Tautenhahn, R., Böttcher, C., and Larson, T. (2012). CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Analytical Chemistry*, **84**(1), 283–289.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software*, **28**(5), 1–26.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., and Engelhardt, A. (2011). *caret: Classification and Regression Training*. R package version 5.05.004.

Kuhn, S., Egert, B., Neumann, S., and Steinbeck, C. (2008). Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction. *BMC bioinformatics*, **9**, 400.

Laursen, K. H., Schjoerring, J. K., Olesen, J. E., Askegaard, M., Halekoh, U., and Husted, S. (2011). Multielemental fingerprinting as a tool for authentication of organic wheat, barley, faba bean, and potato. *Journal of Agricultural and Food Chemistry*, **59**(9), 4385–4396.

Lee, M. (2003). *Statistics::KruskalWallis*. CPAN module version 0.01.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R news*, **2/3**(December), 18–22.

Lommen, A. (2009). MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Analytical chemistry*, **81**(8), 3079–86.

Lommen, A. and Kools, H. J. (2012). MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics*, **8**(4), 719–726.

Mäder, P., Fliessbach, A., Dubois, D., Gunst, L., Fried, P., and Niggli, U. (2002). Soil fertility and biodiversity in organic farming. *Science*, **296**(5573), 1694–7.

Marchini, J. L., Heaton, C., and Ripley, B. D. (2010). *fastICA: FastICA Algorithms to perform ICA and Projection Pursuit*. R package version 1.1-13.

Martens, L., Hermjakob, H., Jones, P., Adamsk, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J. J., and Apweiler, R. (2005). PRIDE: The proteomics identifications database. *Proteomics*, **5**(13), 3537–3545.

Mashego, M. R., Wu, L., Van Dam, J. C., Ras, C., Vinke, J. L., Van Winden, W. a., Van Gulik, W. M., and Heijnen, J. J. (2004). MIRACLE: mass isotopomer ratio analysis of U-13C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites. *Biotechnology and bioengineering*, **85**(6), 620–8.

Matthews, L. and Miller, T. (2000). ASTM Protocols for Analytical Data Interchange. *Journal of Laboratory Automation*, **5**(5), 60–61.

Milgram, E. and Nordström, A. (2009). Metabolomics Survey.

Murray, K. K., Boyd, R. K., Eberlin, M. N., Langley, G. J., Li, L., and Naito, Y. (2013). Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure and Applied Chemistry*, **85**(7), 1515–1609.

Nakayama, K. and Yoshida, H. (1974). PROCESS FOR PRODUCING L-ARGININE BY FERMENTATION.

Neuweger, H. (2009). *MeltDB: a software platform for the analysis and integration of metabolomics experiment data*. Ph.D. thesis, Bielefeld University.

Neuweger, H., Albaum, S. P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J., and Goesmann, A. (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics (Oxford, England)*, **24**(23), 2726–32.

Nielsen, J. and Oliver, S. (2005). The next wave in metabolome analysis. *Trends in biotechnology*, **23**(11), 544–6.

NIST/EPA/NIH (2014). NIST Standard Reference Database 1A.

Nobeli, I., Ponstingl, H., Krissinel, E. B., and Thornton, J. M. (2003). A Structure-based Anatomy of the E.coli Metabolome. *Journal of Molecular Biology*, **334**(4), 697–719.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, **27**(1), 29–34.

Oliver, S. G., Winson, M. K., Kell, D. B., and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in biotechnology*, **16**(9), 373–8.

Orchard, S., Montechi-Palazzi, L., Deutsch, E. W., Binz, P.-A. A., Jones, A. R., Paton, N., Pizarro, A., Creasy, D. M., Wojcik, J., and Hermjakob, H. (2007). Five years of progress in the Standardization of Proteomics Data 4th Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23-25, 2007 Ecole Nationale Supérieure (ENS), Lyon, France. *Proteomics*, **7**(19), 3436–3440.

Park, S. H., Kim, H. U., Kim, T. Y., Park, J. S., Kim, S.-S., and Lee, S. Y. (2014). Metabolic engineering of Corynebacterium glutamicum for L-arginine production. *Nature communications*, **5**, 4618.

Patti, G. J., Yanes, O., and Siuzdak, G. (2012). Innovation: Metabolomics: The apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, **13**(4), 263–269.

Pedrioli, P. G. a., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, **22**(11), 1459–66.

Pence, H. E. H. and Williams, A. (2010). Chemspider: An online chemical information resource. *Journal of Chemical Education*, **87**(11), 1123–1124.

Persicke, M., Rückert, C., Plassmeier, J., Stutz, L. J., Kessler, N., Kalinowski, J., Goesmann, A., and Neuweger, H. (2011). MSEA: Metabolite set enrichment analysis in the MeltDB metabolomics software platform: Metabolic profiling of Corynebacterium glutamicum as an example. *Metabolomics*, **8**(2), 310–322.

Petri, K., Walter, F., Persicke, M., Rückert, C., and Kalinowski, J. (2013). A novel type of N-acetylglutamate synthase is involved in the first step of arginine biosynthesis in Corynebacterium glutamicum. *BMC genomics*, **14**(1), 713.

Pitera, D. J., Paddon, C. J., Newman, J. D., and Keasling, J. D. (2007). Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli. *Metabolic Engineering*, **9**, 193–207.

Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, **11**, 395.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

R_Development_Core_Team (2011). R: A Language and Environment for Statistical

Computing.

Rocca-Serra, P., Salek, R. M., Arita, M., Correa, E., Dayalan, S., Gonzalez-Beltran, A., Ebbels, T., Goodacre, R., Hastings, J., Haug, K., Koulman, A., Nikolski, M., Oresic, M., Sansone, S. A., Schober, D., Smith, J., Steinbeck, C., Viant, M. R., and Neumann, S. (2016). Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics*, **12**(1), 1–13.

Rodgers, R. P., Blumer, E. N., Hendrickson, C. L., and Marshall, A. G. (2000). Stable isotope incorporation triples the upper mass limit for determination of elemental composition by accurate mass measurement. *Journal of the American Society for Mass Spectrometry*, **11**(10), 835–40.

Roessner, U. (2006). The Chemical Challenge of the Metabolome. In D. M. Desiderio and N. M. M. Nibbering, editors, *Metabolome Analysis: An Introduction*, chapter 2, pages 15–38. John Wiley & Sons, Inc., Hoboken.

Röhlig, R. M. and Engel, K. H. (2010). Influence of the input system (Conventional versus organic farming) on metabolite profiles of maize (Zea mays) kernels. *Journal of Agricultural and Food Chemistry*, **58**(5), 3022–3030.

Rojas-Chertó, M., Kasper, P. T., Willighagen, E. L., Vreeken, R. J., Hankemeier, T., and Reijmers, T. H. (2011). Elemental composition determination based on MS(n). *Bioinformatics (Oxford, England)*, **27**(17), 2376–83.

Rojas-Chertó, M., van Vliet, M., Peironcely, J. E., van Doorn, R., Kooyman, M., Beek, T. T., van Driel, M. a., Hankemeier, T., and Reijmers, T. (2012). MetiTree: a web application to organize and process high resolution multi-stage mass spectrometry metabolomics data. *Bioinformatics (Oxford, England)*, pages 2–4.

Royal Society of Chemistry (2014). ChemSpider.

Salek, R. M., Haug, K., Conesa, P., Hastings, J., Williams, M., Mahendraker, T., Maguire, E., González-Beltrán, A. N., Rocca-Serra, P., Sansone, S.-A. A., Steinbeck, C., Gonz??lez-Beltr??n, A. N., Rocca-Serra, P., Sansone, S.-A. A., and Steinbeck, C. (2013a). The MetaboLights repository: curation challenges in metabolomics. *Database : the journal of biological databases and curation*, **2013**, 1–8.

Salek, R. M., Steinbeck, C., Viant, M. R., Goodacre, R., and Dunn, W. B. (2013b). The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience*, **2**(1), 13.

Satman, M. H. (2010). Runiversal: Runiversal - Package for converting R objects to Java variables and XML.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, **36**(8), 1627–1639.

Scheltema, R. a., Jankevics, A., Jansen, R. C., Swertz, M. a., and Breitling, R. (2011). PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass

spectrometry data analysis. *Analytical chemistry*, **83**(7), 2786–93.

Scheubert, K., Hufsky, F., and Böcker, S. (2013). Computational mass spectrometry for small molecules. *Journal of Cheminformatics*, **5**(3), 1–24.

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., Hollender, J., Ru, M., Singer, H. P., and Hollender, J. (2014). Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environmental Science & Technology*, **48**(4), 2097–2098.

Senior, J. K. (1951). Partitions and their representative graphs. *American Journal of Mathematics*, **73**(3), 663–689.

Sévin, D. C., Kuehne, A., Zamboni, N., and Sauer, U. (2015). Biological insights through nontargeted metabolomics. *Current opinion in biotechnology*, **34**, 1–8.

Smedsgaard, J., Villas-Bôas, S. G., Roessner, U., Hansen, M. A. E., Smedsgaard, J., and Nielsen, J. (2006). Analytical Tools. In D. M. Desiderio and N. M. M. Nibbering, editors, *Metabolome Analysis: An Introduction*, chapter 4, pages 83–145. John Wiley & Sons, Inc., Hoboken.

Smith, C., Maille, G., Want, E., Qin, C., Trauger, S., Brandon, T., Custodio, D., Abagyan, R., and Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. *Therapeutic drug monitoring*, **27**(6), 747.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, **78**(3), 779–87.

Stanstrup, J., Neumann, S., and Vrhovsek, U. (2015). PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Analytical Chemistry*, **87**(18), 9421–9428.

Stein, S. E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, **10**(8), 770–781.

Stein, S. E. and Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, **5**(9), 859–866.

Steinbeck, C., Conesa, P., Haug, K., Mahendraker, T., Williams, M., Maguire, E., Rocca-Serra, P., Sansone, S. A., Salek, R. M., and Griffin, J. L. (2012). MetaboLights: Towards a new COSMOS of metabolomics data management. *Metabolomics*, **8**(5), 757–760.

Steinhauser, D., Usadel, B., Luedemann, A., Thimm, O., and Kopka, J. (2004). CSB.DB: a comprehensive systems-biology database. *Bioinformatics (Oxford, England)*, **20**(18), 3647–51.

Sterner, J., Johnston, M., Nicol, G., and Ridge, D. (2000). Signal suppression in electrospray ionization Fourier transform mass spectrometry of multi-component samples. *Journal of mass spectrometry : JMS*, **35**(3), 385–91.

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K. S., Sumner, S., and Subramaniam, S. (2016). Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, **44**(Database issue), D463–D470.

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. a., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., and Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**(3), 211–221.

Sumner, L. W., Lei, Z., Nikolau, B. J., Saito, K., Roessner, U., and Trengove, R. (2014). Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics*, **10**(6), 1047–1049.

Takahashi, H., Morimoto, T., Ogasawara, N., and Kanaya, S. (2011). AMDORAP: Non-targeted metabolic profiling based on high-resolution LC-MS. *BMC bioinformatics*, **12**(1), 259.

Tautenhahn, R., Böttcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, **9**, 504.

Tautenhahn, R., Patti, G. J., Kalisiak, E., Miyamoto, T., Schmidt, M., Lo, F. Y., McBee, J., Baliga, N. S., and Siuzdak, G. (2010). metaXCMS: Second-Order Analysis of Untargeted Metabolomics Data. *Analytical chemistry*, **83**(3), 696–700.

Tautenhahn, R., Patti, G. J., Rinehart, D., and Siuzdak, G. (2012). XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical chemistry*, **84**(11), 5035–9.

Van Der Maaten, L. J. P. and Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, **9**, 2579–2605.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, **10**(5), 988–99.

Villas-Bôas, S., Nielsen, J., Smedsgaard, J., Hansen, M., and Roessner-Tunali, U. (2007). *Metabolome analysis: an introduction*. Wiley, Hoboken, 1st edition.

Villas-Bôas, S. G., Villas-Bôas, S. G., Roessner, U., Hansen, M. A. E., Smedsgaard, J., and Nielsen, J. (2006). Sampling and Sample Preparation. In D. M. Desiderio and N. M. M. Nibbering, editors, *Metabolome Analysis: An Introduction*, chapter 3, pages 39–82. John Wiley & Sons, Inc., Hoboken.

Vitali, R. A., Inamine, E., and Jacob, T. A. (1965). The Isolation of y-L-Glutamyl Peptides from a Fermentation Broth. *The Journal of Biological Chemistry*, **240**(6), 2508–2511.

Vizcaíno, J. A., Côté, R., Reisinger, F., Barsnes, H., Foster, J. M., Rameseder, J., Hermjakob, H., and Martens, L. (2009). The Proteomics Identifications database: 2010 update. *Nucleic Acids Research*, **38**(SUPPL.1), 736–742.

Weckwerth, W. (2003). Metabolomics in Systems Biology. *Annual Review of Plant Biology*, **54**(1), 669–689.

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., Macinnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucleic acids research*, **35**(Database issue), D521–6.

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. a., Lim, E., Sobsey, C. a., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J., and Forsythe, I. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic acids research*, **37**(Database issue), D603–10.

Wittmann, C., Krömer, J. O., Kiefer, P., Binz, T., and Heinzle, E. (2004). Impact of the cold shock phenomenon on quantification of intracellular metabolites in bacteria. *Analytical biochemistry*, **327**(1), 135–139.

Wolf, S., Schmidt, S., Müller-Hannemann, M., and Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, **11**, 148.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, **316**(5827), 1036–9.

Xia, J., Psychogios, N., Young, N., and Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research*, **37**(Web Server issue), W652–60.

Xia, J., Mandal, R., and Sinelnikov, I. (2012). MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis. *Nucleic acids research*, **40**(Web Server issue), W127–133.

Yu, T., Park, Y., Johnson, J. M., and Jones, D. P. (2009). apLCMS–adaptive pro-

cessing of high-resolution LC/MS data. *Bioinformatics (Oxford, England)*, **25**(15), 1930–6.

Zhang, W., Chang, J., Lei, Z., Huhman, D., Sumner, L. W., and Zhao, P. X. (2014). MET-COFEA: A Liquid Chromatography/Mass Spectrometry Data Processing Platform for Metabolite Compound Feature Extraction and Annotation. *Analytical chemistry*, **86**(13), 6245–53.

ERKLÄRUNG

Ich, Nikolas Kessler, erkläre hiermit, dass ich die Dissertation selbständig erarbeitet und keine anderen als die in der Dissertation angegebenen Hilfsmittel verwendet habe.

Bremen, den 9. April 2018

Nikolas Kessler