

# Informatics for Tandem Mass Spectrometry-based Metabolomics



Dipl.-Ing. (FH) Stephan A. Beisken, M.Res.

European Molecular Biology Laboratory

European Bioinformatics Institute

University of Cambridge

Gonville & Caius College

A thesis submitted on April 10, 2014  
for the Degree of Doctor of Philosophy



*“The demand upon a resource tends to expand to match the supply  
of the resource. The reverse is not true.”*

- Generalization of Parkinson's law -



This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee.

April 10, 2014

Stephan A. Beisken



## Acknowledgements

First, I would like to express my gratitude to my supervisor, Dr. Christoph Steinbeck, who has supported me throughout my thesis. His knowledge and life experience helped me to overcome the various hurdles encountered.

I want to thank thank my Thesis Advisory Committee members, Dr. Jeroen Krijgsveld, Dr. Jules Griffin, Dr. John Marioni, and Dr. Mark Seymour, for their advise and guidance. My special thanks goes to Dr. Mark Seymour for providing me with insights into the needs of the industry sector.

I also wish to thank the members of the Steinbeck group and, in particular, the research team. Their assistance and friendship helped me to enjoy my work and push beyond my own limits.

Special thanks to Mr. Mark Earll, Dr. David Portwood, Dr. Reza Salek, and Dr. Michael Eiden for their helpful discussions and suggestions. Their input inspired and helped me to find my way around the data analysis landscape and opened up many glorious opportunities.

The Syngenta AG in collaboration with the European Bioinformatics Institute has provided the funding, support, and data I needed to produce and complete my thesis.

Finally, I wish to thank Ms. Evelyn Lim and my family for their continuous support throughout the programme, ultimately resulting in the creation of this dissertation.





## Abstract

Metabolomics is a rapidly expanding field with applications in areas such as medicine, agriculture, or food safety. Tandem mass spectrometry ( $MS^n$ ) is one of the main technologies that drives the field forward. Optionally coupled to a chromatographic element,  $MS^n$  can capture detailed snapshots of an organism's metabolome. The resulting data sets are complex and difficult to analyse due to the multitude of external, biologically irrelevant influences. In particular metabolite identification – the ultimate goal of  $MS^n$  metabolomics – is a highly challenging exercise with inherently uncertain results.

We have developed the data processing tool *MassCascade* to rapidly analyse and visualise chromatography  $MS^n$  data. *MassCascade* features methods for data (pre-)processing from initial file input to the compilation of the final result matrix. To simplify use and break down the complex analysis process, the tool has been made available in the form of a plug-in for the workflow platform KNIME: *MassCascade-KNIME* offers a visual representation of each processing function that can be utilized following the concept of visual programming. To further support metabolomics data analysis, cheminformatics methods have been added separately to the workflow platform from the Chemistry Development Kit to enable digital small molecule handling, essential for semi-automated metabolite identification.

To demonstrate the  $MS^n$  analysis process and test *MassCascade* and its plug-in, two scenarios typical in metabolomics were chosen: spectral fingerprinting and metabolite identification. A set of metabolomics tomato samples from a long-term study about chromatography  $MS^n$  system stability was processed and interpreted. Distinct trends and clustering could be extracted and explained verifying correct processing by the tool. Metabolite identification of spectral features was applied on a study about tomato ripening. Features differentiating ripening of four different tomato genotypes were singled out to that end. The implemented information-driven identification methodology enabled the selection of putative metabolite identifications from large lists of chemical compounds.



# Contents

Contents	xii
List of Figures	xv
Nomenclature	xviii
<b>1 General Introduction</b>	<b>1</b>
1.1 Metabolomics . . . . .	1
1.1.1 Experimental Methods in Metabolomics . . . . .	4
1.1.2 Applications of Metabolomics . . . . .	5
1.1.3 Mass spectrometry . . . . .	6
1.1.4 Data Pre-Processing . . . . .	13
1.1.5 Data Post-Processing . . . . .	19
1.1.6 Identification of Metabolites . . . . .	20
1.1.7 Software . . . . .	22
1.2 Cheminformatics support for Metabolomics . . . . .	24
1.2.1 Small Molecule Library Management . . . . .	24
1.2.2 Representation of Small Molecules . . . . .	25
1.2.3 Properties of Small Molecules . . . . .	27
1.2.4 Workflow Environments for Cheminformatics . . . . .	28
1.3 Aim of this Thesis . . . . .	33
<b>2 Informatics for LC-MS<sup>n</sup> Analysis</b>	<b>35</b>
2.1 Introduction . . . . .	35
2.2 MassCascade's Implementation . . . . .	36
2.3 MassCascade's Functionality . . . . .	41
2.3.1 Data Pre-Processing . . . . .	41
2.3.2 Data Processing . . . . .	49
2.3.3 Data Post-Processing . . . . .	53
2.4 MassCascade for KNIME . . . . .	57
2.4.1 Structure . . . . .	57

## CONTENTS

---

2.4.2	Node Types . . . . .	59
2.4.3	Node Interactions . . . . .	60
2.5	Evaluation . . . . .	63
2.5.1	Spectral Fingerprinting of Tomato Samples . . . . .	63
2.5.2	Materials and Methods . . . . .	64
2.5.3	Results . . . . .	66
2.5.4	Performance and Scaling of the Core Library . . . . .	78
2.6	Technical Validation . . . . .	80
2.6.1	Methods . . . . .	81
2.6.2	Results & Discussion . . . . .	82
2.7	Conclusion . . . . .	83
2.8	Software Availability . . . . .	83
2.8.1	Update Site . . . . .	84
2.8.2	Extensions . . . . .	84
2.8.3	Example Workflows . . . . .	84
<b>3</b>	<b>Knowledge-based Compound Identification</b>	<b>85</b>
3.1	Introduction . . . . .	85
3.1.1	Open Data . . . . .	87
3.2	Metabolite Identification . . . . .	88
3.2.1	Identification Factors . . . . .	88
3.2.2	Scoring Schemes . . . . .	92
3.3	Materials and Methods . . . . .	93
3.3.1	Tomato Cultivars . . . . .	94
3.3.2	Sample preparation . . . . .	95
3.3.3	Chromatography . . . . .	95
3.3.4	Mass Spectrometry . . . . .	96
3.3.5	Reference Standards . . . . .	96
3.3.6	Data Deposition . . . . .	96
3.4	Data Processing and Transformation . . . . .	97
3.4.1	Known Identification . . . . .	101
3.4.2	Known Unknown Identification . . . . .	102
3.4.3	Unknown Identification . . . . .	102

## CONTENTS

---

3.5	Results . . . . .	103
3.5.1	Analysis of the Quality Controls . . . . .	103
3.5.2	Analysis of the Tomato Samples . . . . .	103
3.5.3	Identification . . . . .	108
3.6	Discussion . . . . .	120
3.7	Conclusion . . . . .	121
3.8	Technical Validation . . . . .	123
3.8.1	Methods . . . . .	123
3.8.2	Results & Discussion . . . . .	124
<b>4</b>	<b>Workflows for Cheminformatics</b>	<b>127</b>
4.1	Introduction . . . . .	127
4.2	KNIME-CDK's Implementation . . . . .	128
4.2.1	Structure . . . . .	129
4.2.2	Persistence . . . . .	129
4.3	KNIME-CDK's Functionality . . . . .	130
4.3.1	Input/Output . . . . .	131
4.3.2	Processing . . . . .	133
4.3.3	Visualisation . . . . .	133
4.4	Evaluation . . . . .	134
4.4.1	Round Tripping . . . . .	138
4.4.2	Test Workflows . . . . .	139
4.4.3	Performance and Scalability . . . . .	139
4.5	Conclusion . . . . .	140
4.6	Software Availability . . . . .	140
4.6.1	Update Site . . . . .	142
4.6.2	Extensions . . . . .	142
4.6.3	Example Workflows . . . . .	142
<b>5</b>	<b>Summary and Discussion</b>	<b>143</b>
	<b>Appendix</b>	<b>147</b>
	<b>References</b>	<b>171</b>



## LIST OF FIGURES

---

1.1	Schematic of a mass spectrometry pipeline and data landscape . . .	7
1.2	Schematic of a mass chromatogram and a spectrum . . . . .	11
1.3	Schematic of the mass spectrometry data analysis process . . . . .	13
1.4	Summary of components contributing to signal distortions . . . . .	16
1.5	Comparison of graphical representations for chlorophyll <i>f</i> . . . . .	27
1.6	Schematic of the principle mechanism of workflow platforms . . . . .	29
1.7	Screenshot of the KNIME workbench . . . . .	31
2.1	MassCascade data types and their representations . . . . .	38
2.2	UML diagram of MassCascade's data structure . . . . .	39
2.3	Example of JavaDoc documentation . . . . .	40
2.4	Comparison of centroiding methods . . . . .	42
2.5	Illustration of noise removal pre-processing methods . . . . .	44
2.6	Illustration of the feature extraction process . . . . .	45
2.7	Illustration of the TopHat algorithm . . . . .	46
2.8	Illustration of Durbin Watson filtering and pre-processing summary	48
2.9	Illustration of deconvolution, alignment, and feature set methods .	51
2.10	Illustration of the modified Bieman algorithm . . . . .	53
2.11	Illustration of annotation and identification methods . . . . .	55
2.12	Linear regression for molecular masses vs. isotope abundances . . .	56
2.13	Schematic of the MassCascade-KNIME architecture and node model	58
2.14	Screenshot of a complex MassCascade-KNIME workflow . . . . .	59
2.15	Schematic of of the node architecture and interactions . . . . .	61
2.16	Screenshot of configuration dialogues . . . . .	62

## LIST OF FIGURES

---

2.17	Screenshot of the <i>Spectrum Viewer</i> data view . . . . .	63
2.18	Schematic of LC-MS data processing for metabolomics fingerprinting	65
2.19	Cross-sample total ion currents and chromatograms . . . . .	67
2.20	Line plot of time vs time deviation of aligned samples . . . . .	68
2.21	Line plot of time vs time drift of aligned samples by group . . . . .	69
2.22	Analysis of interferences intensities and distribution . . . . .	70
2.23	Stepwise analysis of feature missingness . . . . .	72
2.24	Principal component analysis for all standard aliquots . . . . .	75
2.25	Principal component analysis for filtered standard aliquots . . . . .	76
2.26	Analysis of standard aliquots from 2010-09-21 . . . . .	77
2.27	Performance charts of the core library . . . . .	79
2.28	F-scores for feature isolation . . . . .	82
3.1	Workflow for known and known unknown metabolite identification	90
3.2	Processing workflow for metabolite identification . . . . .	98
3.3	Annotated correlation heatmap of tomato study features . . . . .	99
3.4	Overview of the tomato cultivars data set . . . . .	104
3.5	PCA model for the tomato cultivars data set . . . . .	106
3.6	OPLS model for the tomato cultivars data set . . . . .	107
3.7	Pairwise loadings of OPLS genotype models . . . . .	109
3.8	Univariate statistics for features 118.086 and 130.05 . . . . .	117
3.9	Univariate statistics for features 133.061 and 176.103 . . . . .	118
3.10	Univariate statistics for features 197.096 and 327.118 . . . . .	119
4.1	Schematic of the KNIME-CDK architecture and node model . . . . .	129
4.2	Screenshot of a KNIME-CDK workflow . . . . .	131
4.3	Screenshot of KNIME-CDK visualisation preferences . . . . .	134
4.4	Pairwise comparison of measured and calculated logP values . . . . .	137
4.5	Execution times per molecule by different cheminformatics plug-ins	141



## NOMENCLATURE

---

### Symbols

$F_{m/z}$	Feature	$F_{m/z} = (t, I, rt)$
$FS$	Feature Set	$FS_t = \{F_1, F_2, \dots, F_n\}$
$I$	Intensity	
$ma$	Mass Accuracy	
$mcq$	Mass Chromatographic Quality	
$m/z$	Mass-to-Charge Ratio	
$R$	Resolution	$\frac{m/z_1}{\Delta m/z}$
$rt$	Retention time	
$s_i$	Signal	$s_i = (m/z, I)$
$S$	Scan	$S_t = \{s_1, s_2, \dots, s_m\}$
$t$	Time	

## Nomenclature

---

### Acronyms

API	Application Programming Interface
CDK	Chemistry Development Kit
CODA	Component Detection Algorithm
DW	Durbin Watson Criterion
EI	Electron Impact
ESI	Electron Spray Ionisation
FWHM	Full Width at Half Maximum
GC	Gas Chromatography
GCxGC	Two dimensional Gas Chromatography
KNIME	Konstanz Information Miner
LC	Liquid Chromatography
MSI	Metabolomics Standards Initiative
MS	Mass Spectrometry
MS <sup>n</sup>	Tandem Mass Spectrometry
NMR	Nuclear Magnetic Resonance
OPLS	Orthogonal Partial Least Squares
PCA	Principal Component Analysis
PSI	Proteomics Standards Initiative
S/N	Signal to Noise Ratio

## GENERAL INTRODUCTION

---

### 1.1 Metabolomics

Metabolomics is defined as the study of the total small molecule complement of an organism. It is a highly data-generating and knowledge-driven science. The study of the small molecule complement creates a large amount of information-rich data that provides unprecedented insights into an organism's biology within different biological levels such as the tissue, cell type, or compartment level. The metabolome is the dynamic system comprised of the small molecules and their interactions. In the context of an abstract, all-encompassing metabolome, the metabolome can be considered as the ultimate expression of the genome. It provides insights into direct and indirect control and regulation mechanisms of systems. By comparison to other *omics* such as transcriptomics or proteomics, the metabolome is the closest measurable representation of the phenotype currently available, making its potential incalculable<sup>[1]</sup>.

The concept of metabolomics – the word itself is derived from the Greek word for change ( $\mu\epsilon\tau\alpha\beta\omicron\lambda\eta$ ) – was described by C. H. Waddington in 1942: he referred to the study of the causal relationships between genotype and phenotype as *epigenetics*<sup>[2]</sup>.

## 1. GENERAL INTRODUCTION

---

Today, Waddington’s definition of epigenetics describes multiple *omics* disciplines of which metabolomics forms a part of. In contrast to other *omics*, metabolomics has several unique characteristics that make its study particularly demanding: chemical diversity, chemical dynamic range, and time resolution<sup>[3]</sup>.

**Chemical diversity:** Metabolomics studies small molecules within a molecular mass range of 50 to 1500 Da. Chemical classes include amino acids, sugars, alkaloids, phenolic compounds, lipids, and many more. Each class has dramatically different physicochemical properties and biological functions. A simple exchange of a functional group – the smallest functional unit – of a molecular species can change its biological function entirely. A change in stereochemistry can have the same effect. Furthermore, metabolites are not only chemically diverse, they are also hard to enumerate because of the lack of sensible structural and biological constraints.

Wherever enumeration is possible, it is applied. For instance, lipids encompass well defined chemical classes with discrete building blocks. Enumerating biological relevant lipid species is a heavily studied exercise<sup>[4,5]</sup>. Beyond the well defined lipids, *in vivo* phase I and II reactions in addition to other catabolic and anabolic reactions, produce chemical diversity that is challenging to manage<sup>[6]</sup>.

**Chemical dynamic range** refers to the concentration range at which metabolites occur *in vivo*. Depending on the chemical class and location of a metabolite, these can easily span three orders of magnitude or more, e.g. from  $\mu\text{mol/L}$  (hormones) to high  $\text{mmol/L}$  (sugars) concentrations<sup>[7]</sup>. The co-occurrence of metabolites with a 1,000-fold difference in concentration make their simultaneous detection demanding.

**Time resolution** relates to the kinetics and dynamics of metabolites, e.g. the rate at which metabolites degrade over time. Concentrations of metabolites can rapidly change over time. For example, in blood plasma catecholamines have a half-life in the order of minutes whereas the thyroid hormone can have a half-life in the order of hours<sup>[8,9]</sup>. Consequently, any multiparametric responses measured over time can vary significantly from one another, thus constraining the reproducibility of studies.

## 1. GENERAL INTRODUCTION

---

The characteristics outlined above explain why it is inaccurate to talk about *the* metabolome of an organism when referring to discrete biological functions. The notion of a single metabolome is further countered by more recent studies on genome mosaicism<sup>[10]</sup>. Metabolomics experiments acquire snapshots of a metabolome, which properties depend on the sample type. The snapshots can reflect a spatially or temporally constrained aspect of an organism’s state under partially defined conditions.

The aim of metabolomics studies is typically to characterise biological samples or identify metabolites or both based on metabolomics snapshots<sup>[11,12]</sup>. Depending on the study, the set-up can either be targeted (hypothesis-driven) or untargeted (data-driven)<sup>[13]</sup>. The four principal approaches are<sup>[14,15]</sup>:

- fingerprinting, spectral pattern recognition for clustering or identification;
- profiling, description of known chemical classes;
- target analysis, measurement of specific compounds;
- metabolomics, identification of all molecular species in a sample.

The total size of a metabolome is hard to estimate. Any estimation depends on criteria such as the molecular mass cut-off of included metabolites or whether exogenous molecules are included.

Estimates have been attempted for the total number of metabolites for whole kingdoms, e.g. 200,000 for the plant kingdom<sup>[14]</sup>, and for individual organisms, e.g. 9,000 for *homo sapiens*<sup>[16]</sup>. Given our limited understanding of the chemical rules that define observed biological subsets in chemical space, these numbers should be considered with caution. At the moment, all chemical and metabolomics databases do not contain sufficient information to comprehensively retrieve all known metabolites<sup>[17]</sup>. In a recent Nature Review, the authors concluded that “an astounding number of metabolites remain uncharacterized with respect to their structure and function...”<sup>[18]</sup>.

Metabolomics, following in the footsteps of proteomics, is a rapidly growing field<sup>[19]</sup>. The Metabolomics Standards Initiative (MSI)<sup>[20]</sup> was founded in 2007 to address issues related to reporting standards and consolidating community ef-

## 1. GENERAL INTRODUCTION

---

forts – much alike efforts carried out by the Proteomics Standards Initiative (PSI) earlier<sup>[21]</sup>. Experimental studies to characterise different metabolomes on various biological levels have been undertaken, slowly increasing the available knowledge base<sup>[22]</sup>. These efforts have been supplemented by computational approaches for *in silico* metabolite generation and database design to consolidate and stratify collected data on an organism-specific level.

Recent efforts include the development of cross-species metabolomics resources such as MetaboLights<sup>[23]</sup>, the Plant Metabolomics Resource<sup>[24]</sup>, and MeltDB<sup>[25]</sup>. These resources attempt to capture all evidence from a study. This includes, *inter alia*, metabolite structures and their reference spectra, biological roles, locations and concentrations, as well as experimental data. With the aggregation of metabolomics data, the study of data fusion, e.g. from proteomics and metabolomics studies, has gained more attention<sup>[26]</sup>, pushing towards a more integrative and systemic view of analytical sciences.

### 1.1.1 Experimental Methods in Metabolomics

Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR) are the two principal methods used in metabolomics experiments. These highly accurate and sensitive methods are often combined with a chromatographic technique such as High-Performance Liquid Chromatography (HPLC) or Gas Chromatography (GC) to increase spatial resolution (separate molecular species), adding a time dimension to the already complex signal landscape. Sizes of information-dense MS and NMR data range from several to hundreds of Gigabytes.

Chromatography is an important step in metabolomics experiments to separate individual molecular species in a mixture. Advances in chromatographic technology enable the separation of complex mixtures under a variety of experimental conditions<sup>[27,28]</sup>. Denser and more orderly packed columns in combination with higher pressures have the potential to produce sharper signals and shorten run time while maintaining appropriate resolution. In gas chromatography, two dimensional approaches have gained acceptance, yielding unparalleled separation of complex mixtures<sup>[29]</sup>.

## 1. GENERAL INTRODUCTION

---

NMR and MS technologies are complementary, detecting different chemical classes or molecular species. In general, NMR requires less time-consuming sample preparation, is reproducible, and quantitative. MS is more sensitive and high-throughput in comparison<sup>[30]</sup>. These technologies have been shown to produce similar results for high-level applications such as fingerprinting<sup>[31]</sup>. Combinations of technologies result in a great number of systems each suited for individual studies. For a review, please see Aliferis and Shulaev *et al.*<sup>[13,32]</sup>.

Here, we focus on liquid chromatography coupled to tandem mass spectrometry (LC-MS<sup>n</sup>), a routine technique used to investigate the small molecule complement of organisms. Modern LC-MS<sup>n</sup> systems can detect more mass traces than ever before thanks to high mass accuracy ( $ma \leq 2$  ppm<sup>[33]</sup>) and high resolution ( $R \geq 100,000$ <sup>[34]</sup>), producing complex, information-rich data for every sample. LC-MS<sup>n</sup> has been applied across different fields in biology<sup>[11,35]</sup>. The diverse variety of available instrumental platforms and configurations<sup>[36,37]</sup> reflect that no single platform or method can cover the whole metabolome<sup>[12]</sup>.

Nevertheless, LC-MS<sup>n</sup> can be applied to the study of lipids and core metabolism in combination with different approaches<sup>[38]</sup>. In environmental science<sup>[39]</sup> and plant science<sup>[40]</sup>, non-targeted metabolomics have been predicted to become of particular importance owing to LC-MS<sup>n</sup>'s ability to resolve a huge range of semi-polar compounds. The basics for the measurement of small molecules, are captured in best practice guides such as published by Webb *et al.*<sup>[41]</sup>. Efforts on quantitative MS are mostly limited to GC-MS due to higher reproducibility of results, which is important for studies involving instrument calibration<sup>[42]</sup>.

### 1.1.2 Applications of Metabolomics

Metabolomics has been applied to a wide variety of areas. For example, mass spectrometry methods have been used in medical diagnostics<sup>[43]</sup>, studies about cancer<sup>[44,45]</sup> and neurological disease<sup>[46]</sup>. Fluids commonly studied by metabolomics in the context of medical studies include urine<sup>[47]</sup>, plasma (serum)<sup>[48,49]</sup>, and cerebrospinal fluid<sup>[50]</sup>. Lipidomics – part of metabolomics but due to its complexity considered a separate field – has drawn attention from the pharma-

## 1. GENERAL INTRODUCTION

---

ceutical sector because of its relevance to diseases like diabetes or obesity<sup>[51,52]</sup>. Metabolomics's non-invasive nature and extremely high time resolution makes it an ideal tool for the pharmaceutical industry.

Metabolomics has also been used in the characterisation and identification of bacterial strains<sup>[53]</sup>, serving as an early-detection system in clinical environments<sup>[54]</sup>. In addition to these specific applications, metabolomics is used in systems biology as one of the many *omics* disciplines that this field tries to combine<sup>[11]</sup>.

Plant metabolomics is particularly interesting because of the range and functions of primary and secondary metabolites in plants<sup>[55]</sup>. About 300 distinct metabolites could be routinely identified a decade ago, a number that has not changed much over time<sup>[56]</sup>. Applications of plant metabolomics include basic research (untargeted approaches<sup>[57,58]</sup>), environmental studies<sup>[59]</sup>, targeted studies<sup>[60]</sup>, profiling of varieties of cultivars<sup>[61,62]</sup>, plant lipidomics<sup>[63]</sup>, and untargeted chemical identification of plants<sup>[64]</sup>.

### 1.1.3 Mass spectrometry

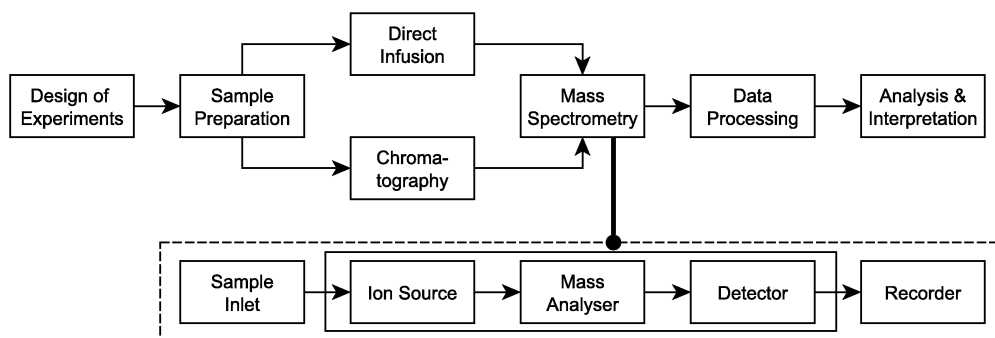
Mass spectrometry is an analytical technique to measure small molecules, either directly injected into the MS or via an interfaced chromatographic technology. The analytes are ionised at an ion source before they can be detected in a coupled mass detector. The resulting data consists of mass-to-charge ( $m/z$ ), time, and intensity triplets that describe for every detected ion mass the strength of the ion beam and the time it is detected (Figure 1.1).

The most common chromatographic technologies used in mass spectrometry are gas and liquid chromatography, distinguished by the state of their mobile phase. These technologies are not as high throughput as direct infusion techniques but suffer less from ion suppression and unresolved isobaric compounds<sup>[65]</sup>. Chromatography adds an additional dimension to the MS data landscape. Through interactions of analytes with a mobile and stationary phase, compounds are retarded and elute off a chromatographic column at different time points due to their physicochemical properties. This allows isobaric species to be resolved.

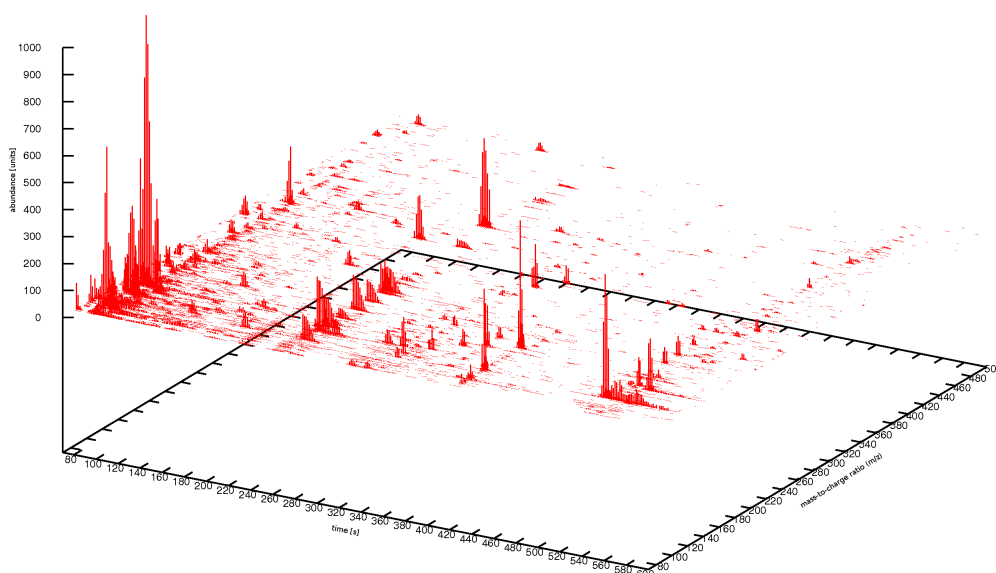


## 1. GENERAL INTRODUCTION

---



(a) Flow diagram of a typical mass spectrometry pipeline



(b) Three dimensional mass spectrometry data landscape

Figure 1.1: Schematic of a typical mass spectrometry pipeline and three dimensional data landscape. (a) Flow diagram of a typical mass spectrometry pipeline from design of experiments to the final interpretation of results. A mass spectrometer can be interfaced with a chromatographic technique or used via direct infusion. The dotted rectangle shows the building blocks of a mass spectrometer. The central parts, ion source, mass analyser, and detector, are a separate unit under high vacuum. (b) Data landscape of chromatography-interfaced MS data. The detector scans over a mass range at discrete time intervals, picking up mass-to-charge ratio ( $m/z$ ) of ions arriving at the detector.

## 1. GENERAL INTRODUCTION

---

Ionisation techniques are grouped into hard and soft. Hard ionisation such as electron impact ionisation (EI), heavily fragments a compound by creating high energy electrons that interact with an analyte. In contrast, soft ionisation techniques, such as electron spray ionisation (ESI), ionise a compound but create only few fragments, for example, based on the principle of Coulomb repulsion. Those techniques can be used separately or in combination<sup>[66]</sup>.

Generated ions are separated by their mass-to-charge ratio ( $m/z$ ) in the mass analyser. For simplicity charge is often assumed to be equal to one. Consequently a mass-to-charge ratio approximately equals the molecular mass of an ion. All mass analysers exploit the mass and electrical charge properties of ions but use different separation methods and vary in performance<sup>[67]</sup>. Finally, separated ions are captured by a mass detector that scans a pre-defined mass range at close intervals. The chromatographic profile of an ion, i.e. the generated continuous ion beam, is captured across multiple scans at discrete time intervals. For a review of LC-MS technologies in metabolomics, see Forcisi<sup>[68]</sup> and Draper *et al.*<sup>[69]</sup>.

Mass spectrometers can be operated in tandem with two (MS/MS) or more (MS<sup>n</sup>) spectrometers working in sequence, fragmenting selected ions further in collision chambers in between individual mass spectrometers. Ions are selected for fragmentation in a data-dependent manner based on the scan mode, e.g. parent ion scan or product ion scan<sup>[70]</sup>. The resulting data does not vary in its structure but has a greater depth. Parent ions from MS<sup>1</sup> have associated scans in MS<sup>2</sup>, MS<sup>3</sup>, *et cetera*. In addition, mass spectrometers can run in positive and negative ion mode, where the mass analyser filters for positive and negative ions respectively. Compounds show different fragmentation patterns for each ion mode. Instruments in positive ion mode have been shown to create more fragments than machines run in negative ion mode<sup>[47]</sup>.

The resulting partially convoluted, densely populated signal landscape contains systematic and random noise amongst true signals of varying intensity and shape. Due to fragmentation and the inevitable presence of contaminants and interferences, compounds are represented by many signals<sup>[71]</sup>. The following provides a breakdown of different signal sources other than fragmentation stemming from the same compound. With soft ionisation techniques, *main ions* are formed through

## 1. GENERAL INTRODUCTION

---

addition or loss of a hydrogen ( $[M+H]^+$  or  $[M-H]^-$ ). *Adducts* can form through interaction with other molecular species such as sodium:  $[M+Na]^+$ . *Clusters* result from aggregation of the compound under investigation with itself:  $[2M+H]^+$ . In addition, charge is not restricted to one. Species with higher charges such as  $[M+2H]^{2+}$  can be observed.

### Properties of Mass Spectrometers

The type and configuration of mass spectrometers dramatically influence the quality of the resulting data in all three dimensions: mass-to-charge ( $m/z$ ), time, and intensity. This section introduces common terms that describe instrumental parameters and characteristics of data acquired by mass spectrometers coupled to a chromatographic method. Depending on the quality of the data landscape, data processing and analysis parameters have to be adjusted, e.g. to account for poor resolution in the  $m/z$  dimension. Therefore, it is essential to understand these descriptors. For an overview and in-depth summary of terms relating to mass spectrometers please see Moco<sup>[36]</sup> and Price *et al.*<sup>[72]</sup>.

A chromatographic component adds a time domain to the signal landscape, which increases the resolution of isobaric compounds. Peaks in chromatograms ideally follow a Gaussian distribution. Chromatograms of a single ion as detected by a mass spectrometer are also known as mass chromatograms or extracted ion chromatograms (Figure 1.2a). They are defined by a characteristic retention time (rt), measured at the apex of the Gaussian-distributed peak, and a maximum peak height. If one chemical species elutes at two different time points, the second peak is referred to as *shadow peak*. If two compounds of similar or identical mass elute at a similar time point, the two chromatograms are said to be convoluted, i.e. they overlap. The two compounds can either be resolved via peak picking (deconvolution) on the data processing side or by increased mass or chromatographic resolution on the instrumental side. In addition to increased resolution, chromatography also reduces ion suppression in the ion source<sup>[73]</sup>. Ion suppression prevents low abundance species to get ionised. Consequently, these species cannot be detected. For a review on ion suppression, please see Furey<sup>[74]</sup>

## 1. GENERAL INTRODUCTION

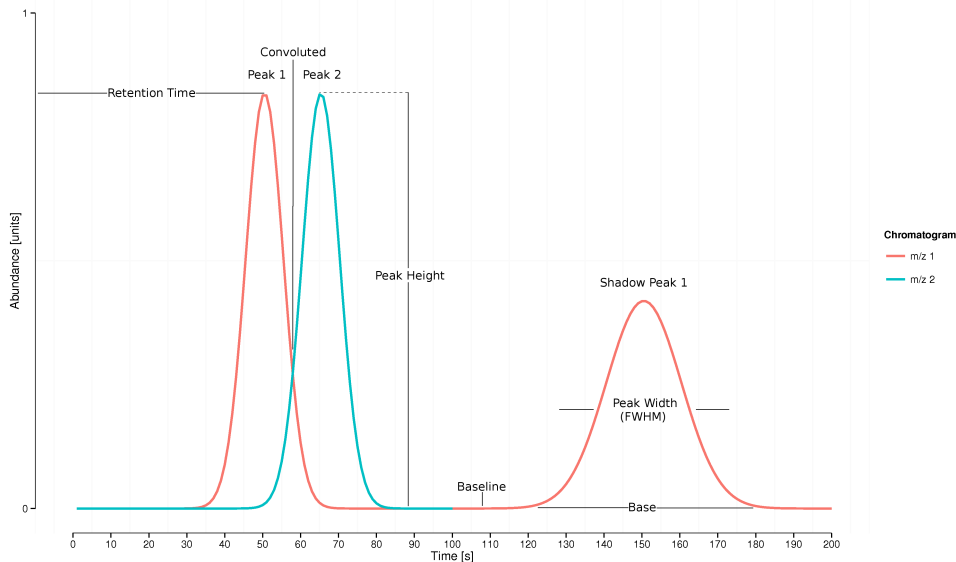
---

and Annesley *et al.*<sup>[75]</sup>. Sensitivity refers to the change in ion current for a compound against the background. It is described by the signal-to-noise ratio (S/N). Higher sensitivity enables the detection of more signals from compounds of lower concentration and less strict background filtering should be considered for data processing. As established previously, mass detectors scan a given mass range at discrete intervals. These intervals are defined by the *scan rate*. Higher scan rates yield more data points per chromatographic signal.

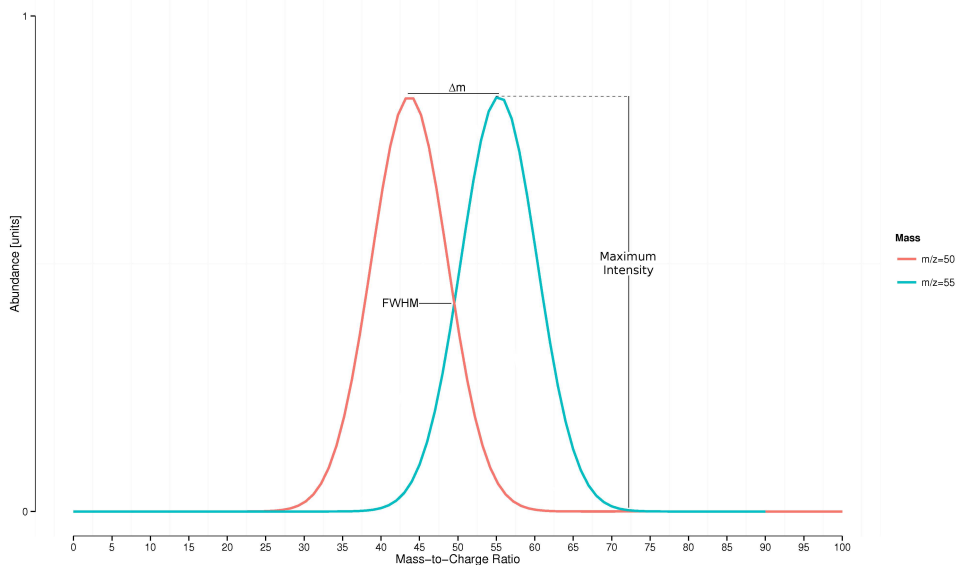
Narrow signals, e.g. chromatographic traces that consist of less than four data points, carry less significance than signals with more data points that follow a well-behaved Gaussian shape<sup>[76]</sup>. Scan rate inversely affects *mass resolution* or *mass resolving power*. Both terms refer to the ability to distinguish two overlapping signals. Following IUPAC's Gold Book recommendations<sup>[77]</sup>, the mass resolving power  $R$  of two overlapping signals is defined as  $R = \frac{m/z_1}{\Delta m/z}$ , where  $m/z$  is the mass of the indexed signal and  $\Delta m/z$  equals  $m/z_1 - m/z_2$ . The extent to which the signals overlap must be indicated by either a percentage (10%) or by FWHM (full-width-at-half-maximum, 50%) of the signal height where the overlap occurs (Figure 1.2b). Closely related, *mass accuracy* describes how precisely a known mass ( $m$ ) can be measured. Deviations from the exact value are specified in parts per million (ppm)<sup>[78]</sup>. Consequently, mass tolerance or mass error, i.e. the allowed  $m/z$  wobble of an ion trace over time, is typically defined in ppm. Mass accuracy of instruments decreases with increasing mass. The parts-per notation is particularly useful because it describes a dimensionless fraction. A mass tolerance of 10 ppm gives 0.001 Da tolerance for a compound of mass 100 Da and 0.008 Da tolerance for a compound of mass 800 Da.

For data processing and exchange, it is convenient to collapse signals of multiple scans into a single spectrum. A spectrum refers to a collection of signals that can originate from multiple scans. Here, the technically more correct term scan will be used interchangeably with the term spectrum.

# 1. GENERAL INTRODUCTION



(a) Schematic mass chromatogram



(b) Schematic mass spectrum

Figure 1.2: Schematic of a mass chromatogram and a spectrum. (a) Example chromatogram of two partially overlapping ion species. Each species has a characteristic retention time (rt) and peak height measured from the baseline. Subsequent peaks of the same ion ( $m/z$  1) are referred to as shadow peaks. (b) Example spectrum of two overlapping  $m/z$  signals. Mass resolution ( $R$ ) can be measured at full-width-at-half-maximum (FWHM) via  $R = \frac{m/z_1}{\Delta m/z}$ , where  $\Delta m/z = m/z_1 - m/z_2$ .

## 1. GENERAL INTRODUCTION

---

### Trends in Mass Spectrometry Metabolomics

A diverse array of mass spectrometers exist, each with unique advantages<sup>[67]</sup>. This section outlines current trends in instrumentation. In-depth reviews and comparisons of existing platforms can be found in the literature<sup>[79,80]</sup>.

Almost all properties of mass spectrometers have improved over the last decade, including mass accuracy, scan rate, and resolution<sup>[65,81]</sup>. Two dimensional gas chromatography (GCxGC) has continued to grow in popularity over recent years, offering increased separation capacity and thus selectivity<sup>[39,82]</sup>. With the fundamental issues addressed, the field is moving into tandem mass spectrometry, catered for by MS vendors<sup>[83,84]</sup>.

While instrumental hardware is constantly improving, mass spectrometry-based metabolomics is lagging behind in comparison to Proteomics with regard to software<sup>[85]</sup> and analysis standards<sup>[86]</sup>, which are only slowly emerging. Most notably, improved instrumentation has enabled advances in untargeted metabolomics<sup>[87,88]</sup> and quantification<sup>[89]</sup>. Cross-sample retention time stability and analyte ionisation paired with high resolution has simplified calibration procedures and increased system stability.

Tandem mass spectrometry refers to the combination of mass spectrometers in sequence ( $MS^n$ ). Selected ions from one mass spectrometer are fragmented further in the next mass spectrometer through a collision chamber. The number of spectrometers in sequence ( $n$ ) is limited by the increasing engineering complexity and diminishing signal, i.e. ion concentration, with every appended mass spectrometer<sup>[87]</sup>. Proof-of-principle studies have employed systems of up to  $MS$  level four<sup>[90]</sup>. Four common data-driven methods for ion selection exist that allow study-based control over  $MS^n$  spectra generation, where all  $MS^n$  spectra follow the precursor/fragment relationship. The precursor isolation window determines the purity of the detected  $MS^n$  spectra. Narrow isolation windows reduce contaminating and interfering ions through increased selectivity but also remove relevant information like isotope patterns<sup>[70]</sup>.

## 1. GENERAL INTRODUCTION

---

Tandem mass spectrometry is gaining popularity for the elucidation of unknown compounds and in data-driven untargeted metabolomics. However, decreasing mass accuracy of MS<sup>n</sup> levels greater than one and complex fragmentation behaviour of small molecules make the interpretation of MS<sup>n</sup> spectra difficult and computationally expensive<sup>[91]</sup>. To complement these technological and computational advances, standard reference materials are under development to facilitate efforts in metabolite identification and quantification<sup>[92]</sup>.

### 1.1.4 Data Pre-Processing

LC-MS<sup>n</sup> data processing includes many steps, most of which modify or remove raw data. Consequently, it is important to establish a good understanding of the steps involved<sup>[93]</sup>. The endpoint of mass spectrometry-based metabolomics studies is an annotated feature matrix extracted from a set of samples (raw data). A feature is defined as signal ( $m/z$ , intensity value pair) that is believed to represent an ion. Multiple features that represent different ions can belong to the same compound due to fragmentation, different ionisation states, adduct formation, or clustering. In contrast, signals originating from noise are not considered features.

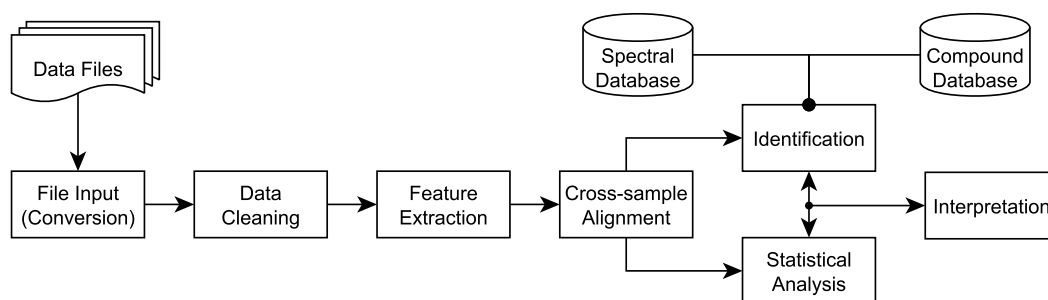


Figure 1.3: Schematic of the mass spectrometry data analysis process. A set of raw data files is read after file conversion to non-proprietary formats. Data cleaning prepares raw data for feature extraction through noise reduction and background correction. Feature extraction isolates ion traces from raw data that are believed to represent a compound, before cross-sample alignment is carried out to compile a feature matrix for statistical analysis. Additionally, features can be identified using spectral and chemical compound databases.

## 1. GENERAL INTRODUCTION

---

To compile the feature matrix, noise reduction and background correction are essential before feature extraction, which greatly clean up the data. Extracted features of individual samples are then aligned across samples to compensate for retention time drifts introduced by the chromatographic component (Figure 1.3). Following, aligned features can be aggregated in a feature matrix, where a feature has a characteristic mass used as column header and the samples represent row identifiers.

### File Formats and Conversion

Mass spectrometry data is stored in a file-based manner where one file typically represents one MS run. Vendor software that operate MS instruments use proprietary file formats that are rarely supported by non-proprietary software tools. Exceptions may occur when (a) the vendor offers an intelligible application programming interface and (b) implementation is easy. In any case, closed proprietary formats impede data exchange and isolate tools that can only implement a limited number of those formats.

Open file formats such as *mzXML*<sup>[94]</sup> and *mzData*<sup>[95]</sup> were developed to address this issue. Originally developed for proteomics, metabolomics has adopted these markup-based standards. The newer HUPO PSI *mzML* 1.1.0<sup>[96]</sup> format has become the *de facto* standard superseding the older formats<sup>[97]</sup>. Notably, *netCDF*<sup>[98]</sup>, a generic common data format, is still used in MS. For an update on the efforts of the HUPO PSI, please see Orchard *et al.*<sup>[99]</sup>.

File format conversion tools bridge the gap between closed proprietary and open formats, partially relying on vendor libraries for accurate conversion. They allow software developed for MS to ignore the plethora of vendor formats by taking over the responsibility of format conversion. This is facilitated by the accepted open data standards outlined above<sup>[100]</sup>.



## 1. GENERAL INTRODUCTION

---

### Data Cleaning

Data cleaning is important to remove irrelevant signals and reduce data size. It includes an array of processes that manipulate raw data that should be applied with care. Baseline drift is a common problem in LC-MS<sup>n</sup> where the gradient of the mobile phase causes the chromatographic baseline to be trending up- or downwards. This complicates analysis because of the baseline's effect on chromatographic peak shapes, introducing fronting or tailing. Distorted peak shapes complicate peak detection and feature extraction (Figure 1.4). Background correction methods have been developed to address this problem<sup>[101–105]</sup>. These algorithms reduce systematic background drift by subtracting either a reference or an estimated background intensity value from the sample chromatogram.

Background correction methods account for systematic errors in the data but do not remove random noise. Random noise produces signal spikes and discontinuous data that could be mistaken for meaningful data. In order to distinguish random noise from meaningful signals, criteria have been developed to evaluate chromatographic signal traces. These include the Component Detection Algorithm (CODA) that measures the mass chromatographic quality (MCQ)<sup>[106]</sup> and the Durbin-Watson (DW) criterion that quantifies randomness<sup>[107]</sup>. Chromatographic traces above a given threshold are considered noise and are removed. Data smoothing forms part of the noise removal process. Smoothing algorithms remove spikes from traces, for example by polynomial regression<sup>[108,109]</sup>. Peak smoothing simplifies feature detection and extraction by modelling chromatographic traces into ideal shapes, smoothing algorithms can also mask noise by modelling random signals into real ones.

Additional data cleaning operations include simple  $m/z$ , time, and intensity filter, which crop raw data and remove irrelevant parts of the data landscape. Removal of traces of known contaminants and interfering ions, such as acetonitrile and methanol products, is also used to remove background noise<sup>[110]</sup>.

## 1. GENERAL INTRODUCTION

---

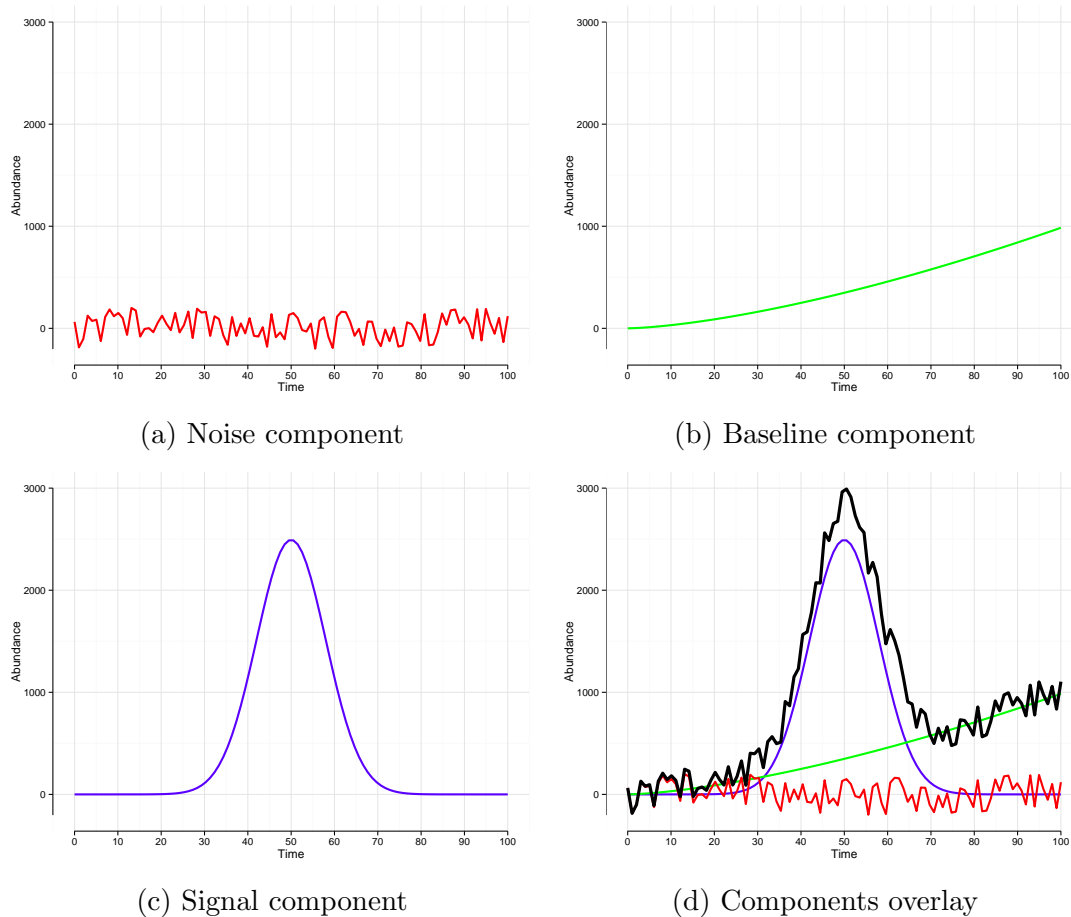


Figure 1.4: Summary of components contributing to signal distortions. (a) Random noise adds variation to a signal around mean zero. (b) Systematic noise, e.g. baseline drifts, introduces a systematic drift or bias in the data that needs to be removed before data analysis. Systematic noise can impact heavily on signal intensities and derived signal areas. (c) The actual signal follows – in theory – a Gaussian distribution. Deviations from this distribution reflect external factors. (d) Overlay of components (a), (b), and (c), and the resulting “measured” signal (black).

## 1. GENERAL INTRODUCTION

---

### Feature Detection

Feature detection and deconvolution describe the process of isolating chromatographic traces of individual ions and splitting these traces into separate peaks<sup>[111]</sup>. A trace is the chromatographic profile of a single ion. A single chromatographic trace with multiple peaks can result from a single compound – eluting off the column at different time points due to matrix effects – or from multiple compounds. Hence, peaks in the same trace need to be distinguished in case they overlap through deconvolution. For a review on feature detection algorithms, please see Zhang *et al.*<sup>[112]</sup>.

Many methods for feature detection of varying complexity have been published. These range from simple procedural approaches<sup>[113–115]</sup> to model-based<sup>[116]</sup> and more abstract approaches using signal segmentation<sup>[117]</sup> or self-modelling curve resolution<sup>[118]</sup>. Routinely applied tools use simple detection methods because of their robustness, speed, and ease-of-use (see section 1.1.7). For a single feature, limitations of a mass detector to reduce the  $m/z$  measurement error to zero, i.e. a mass accuracy of zero ppm, result in a range of detected  $m/z$  values for multiple scans. Because higher intensity signals yield better mass accuracy than lower intensity signals due to instrumental limitations, detected  $m/z$  values are intensity-weighted to determine the most precise mass-to-charge ratio of a feature. A feature’s representative retention time and intensity are then taken from its apex. A  $m/z$  search window – defined by a mass tolerance – is typically described in parts-per-million to define the maximum allowed  $m/z$  deviation of a trace.

Overlapping chromatographic traces, i.e. traces of individual ions that are close or below the mass resolution, need to be flagged. The flagged traces need to be separated by assigning individual data points to the most likely feature or via deconvolution if the traces are indistinguishable. Deconvolution can either be a separate step or part of the feature detection step. Algorithms working on the shape of chromatographic traces try to identify individual features either by finding local maxima<sup>[113]</sup> or by modelling and fitting (ideal) peak shapes<sup>[119]</sup>.

## 1. GENERAL INTRODUCTION

---

### Sample Alignment

Retention times of compounds vary from sample to sample due to matrix effects, altered column conditions, pressure differences, and additional technical limitations. These retention time drifts can range from a few to several seconds and pose a major obstacle for cross-sample comparisons of features<sup>[120]</sup>. Experimentally, retention time drifts can be reduced through column conditioning. Initial column conditioning and between-run column equilibration to the original conditions ensure that column performance remains as constant as possible.

Computationally, algorithms for *time warping* have predominantly been developed for spectroscopy applications in general. However, the same algorithms can be used for metabolomics LC-MS data. They work on either raw data or extracted features and group signals/features across samples by correlation, accounting for the non-linear nature of retention time deviations. Existing methods are based on time warping<sup>[121–123]</sup>, clustering followed by time corrections<sup>[124,125]</sup>, or variance-based approaches<sup>[126]</sup>.

### Spectrum Extraction

An extracted spectrum consists of a set of correlated signals. In the ideal case, all signals result from the same molecular species captured, i.e. the spectrum may contain signals from fragments and adducts as well as ion clusters. Such a spectrum is called a compound spectrum or feature set and can be used in identification<sup>[127]</sup>.

The primary criterion for correlation is retention time. Ions that arrive at the detector simultaneously have either eluted off the chromatographic column together or have formed during the ionisation process. For high chromatographic resolution, these sets of signals result only from few molecular species. Consequently, the dominant approach to spectrum extraction is the aggregation of signals across individual scans around a given retention time.

More elaborate methods use additional criteria such as the shape of an ion chromatogram. Signals are only grouped together into a compound spectrum if the

## 1. GENERAL INTRODUCTION

---

retention time and the elution profile are similar. This enables separation of co-eluting compounds. Adduct information can also be used for correlation. These methods yield cleaner, less noisy, compound spectra<sup>[128,129]</sup>.

### 1.1.5 Data Post-Processing

Data post-processing refers to the statistical analysis and interpretation of processed data. Extracted and aligned features (or compound spectra) can be collected in a feature matrix, where a feature has a characteristic mass used as column header and the samples represent row identifiers. The values at the sample-feature intersections are intensities. Analysis of the matrix includes data normalization and annotation of related features and, ultimately, interpretation of the results<sup>[130,131]</sup>. Feature annotation includes identification as discussed in the following subsections.

#### Statistical Analysis

Statistical analysis methods can be grouped into univariate and multivariate, each offering unique insights into the data. Multivariate analysis works on a matrix of variables. It highlights characteristics based on the relationships between all variables. Univariate analysis takes only one variable into account, resulting in differently weighted results.

The goal of statistical analysis is the categorisation and prediction of sample properties through generation of models that capture the information contained in data matrices. In mass spectrometry, the  $m/z$ -ratio and signal intensity are the two most important variables<sup>[132]</sup>.

Without venturing far into the area of *Chemometrics*, principal component analysis (PCA) and (orthogonal) partial least squares (PLS) are established methods for multivariate analysis of mass spectrometry data. These methods extract latent variables by maximum variance and maximum covariance to the dependent variable respectively. The dimensionality-reduction methods can be used in classification, regression, and prediction exercises<sup>[133,134]</sup>. The quality of statistical

## 1. GENERAL INTRODUCTION

---

models built from the data depend significantly on data pre-processing as well as scaling and normalization. This requires careful investigation of multiple models for consensus building<sup>[135,136]</sup>.

### 1.1.6 Identification of Metabolites

Metabolite identification of signals or compound spectra is an important goal of metabolomics mass spectrometry experiments. Identified metabolites yield in-depth biological insight in addition to information retrieved from spectral fingerprinting. The challenge of metabolite identification lies in the vast tangible chemical space and limitations in available reference data. As little as 10% of extracted features may be of true biological origin<sup>[110]</sup>, where non-biological features result from adduct formation, clustering, interferences, and noise. The available chemical solution space covers most of those irrelevant features, which increases the chance of false identifications. Even for signals of biological origin, multiple identification results are feasible for a single feature, complicating the ranking of these results. The possibility to narrow down the solution space through experimental reference data is hampered by limited numbers of reference data and by issues related to cross-comparisons of reference spectra from different instruments or methods.

#### The Identification Process

The identification process starts from features and compound spectra that are queried against databases that contain relevant metabolites and reference spectra. In case of a single feature, a characteristic  $m/z$ -value is used as query criterion for which, within a given mass tolerance, matching chemical structures are retrieved. Stereoisomers cannot be resolved by mass spectrometry alone because identification methods are mass based. For compound spectra, spectra queries provide a powerful and less generic way to retrieve putative metabolite identifications. Instead of querying a single  $m/z$ -value, the complete spectral vector that characterizes a metabolite is used for the search. Spectra queries depend

## 1. GENERAL INTRODUCTION

---

on databases that contain reference spectra from identical or similar instruments with similar configurations to be reliable, dramatically reducing the available query space. The problem of reference data is more relevant for LC-based than GC-based metabolomics because of the more consistent GC retention time and GC-MS fragmentation pattern<sup>[137]</sup>. Thus, rich databases are corner stones for metabolite identification<sup>[138]</sup>. Queries can return zero to many results – possibly already ranked by similarity – that need to be re-ranked on additional information and interpreted in biological context before a single compound can confidently be chosen as identity for the query feature or compound spectrum. Additional information include fragmentation spectra, isotope patterns, or orthogonal information such as time-of-flight or retention time. These help to narrow down a list of potential metabolite identifications based on molecule specific properties. Biological information, e.g. through utilization of pathway maps or modelling, provides the necessary context to increase the confidence in identifications.

### Reporting Standards

Capturing the minimum set of information to reproduce a metabolomics study is of paramount importance to simplify data exchange and ultimately guarantee good quality of work. Reporting standards outlining the information required for particular technologies such as MS<sup>[139]</sup> or NMR<sup>[140]</sup> are under development. These frameworks need to be adopted by the community and consumed by software tools to be effective<sup>[141]</sup>. To this end, the *mzML* file format has already started to replace older file formats and the *mzTab* file format has been developed for the reporting of identification results. The *mzTab* file format is still undergoing review within PSI at the time of this writing. In parallel, an increasing number of software tools support the new file formats and projects have been launched to collect metabolomics data adhering to minimum reporting standards and to harmonise existing standards further<sup>[142,143]</sup>.

## 1. GENERAL INTRODUCTION

---

### 1.1.7 Software

A variety of software tools have been developed for MS<sup>n</sup> data processing and analysis. Given the complexity of the task, the majority of released software packages focus on individual steps, e.g. feature alignment or noise reduction, and some offer an all-in-one approach. These include processing algorithms as well methods for metabolite identification or statistical analysis (Table 1.1).

However, even all-in-one tools cannot offer all of the functionality needed because of the heterogeneous nature of metabolomics data and unforeseeable advances in the field. Consequently, pipelines concatenating existing tools are constantly being built<sup>[171-174]</sup>. These, in turn, act as guide for the development of the next generation of expert all-in-one tools.

Both proprietary and free software libraries can be grouped into three categories: command-line, stand-alone graphical user interface (GUI), and web-based tools. Each offering unique advantages, frequently reviewed and discussed in literature<sup>[70,175,176]</sup>. A further distinction must be made with regard to the chromatographic method being used. Independent of the actual experimental method, data properties vary for different instruments. Subsequently most tools are optimized for either gas or liquid chromatography, or are even more specific for one technology, e.g. capillary electrochromatography or time-of-flight mass spectrometry. This, in combination with the continuous increase in mass accuracy and throughput of modern machines<sup>[177]</sup>, also acts as driver for software development in MS. Leaps in technology, such as the advent of two dimensional gas chromatography instruments (GCxGC), followed by the rise of GCxGC software, e.g. for alignment<sup>[178]</sup>, illustrate this point nicely<sup>[179,180]</sup>.



# 1. GENERAL INTRODUCTION

CLI			GUI			Web		
Name	Cited	Ref.	Name	Cited	Ref.	Name	Cited	Ref.
<i>All-in-One</i>								
Metab	23	[144]	AMDIS	386	[113]	MetaboAnalyst	257	[145]
MetSign	14	[146]	MzMine2	272	[147]	TOPSIMS-P	3	[148]
PyMS	9	[149]	OpenMS	227	[150]			
eMZed	2	[151]	Met-IDEA	154	[152]			
			MetaboliteDetector	77	[153]			
			MAVEN	63	[154]			
			mzMatch	46	[155]			
			TracMass2	1	[156]			
			MAIT	0	[157]			
<i>Processing</i>								
XCMS	974	[158]	MetAlign	240	[159]	XCMSWeb	85	[160]
TargetSearch	39	[161]	PrepMS	40	[162]			
AMDORAP	4	[163]	MaSDA	2	[164]			
X13 CMS	1	[165]	MSeasy	2	[166]			
<i>Identification</i>								
MolFind	11	[167]				MZedDB	67	[168]
						MetFusion	19	[169]
						MetiTree	4	[170]

Table 1.1: Table of non-commercial mass spectrometry software for metabolomics. The tools have been divided column-wise in command line (CLI), graphical user interface (GUI), and website-based (Web) tools. The All-in-One group includes software that offers functionality for data processing and analysis or identification. The tools listed in the Processing and Identification group focus exclusively on data (pre-)processing and signal identification respectively. Individual groups are sorted in descending order by the number of citations. The citation numbers ('Cited') for the referenced articles are taken from Google Scholar. No distinction has been made between software for MS in tandem with gas or liquid chromatography.

### 1.2 Cheminformatics support for Metabolomics

Increasing computational power has enabled the rise of cheminformatics<sup>[181]</sup>. The principles of cheminformatics – the fusion of computer science and chemistry – were first described in the 1970s and 1980s, but only attracted wide recognition with the dawn of powerful personal computers a couple of decades ago<sup>[182]</sup>. Similar to bioinformatics, cheminformatics has developed into a separate field of study that penetrates into many areas of modern life science<sup>[183]</sup>.

Data-driven metabolomics inherently depends on cheminformatics<sup>[184]</sup>. Experimental methods generate information-rich data that at their core describe molecular structures. For example, chemistry databanks such as PubChem<sup>[185]</sup> and ChemSpider<sup>[186]</sup>, with over 47 million and 29 million structures respectively, are back-ends for metabolomics applications. Querying those databases in a semi- or fully-automated fashion and analysing the results is at the very core of cheminformatics. Model building for clustering, prediction of chemicals or chemical properties, and pathways modelling for biological interpretation are further use cases of cheminformatics in metabolomics. Cheminformatics tool kits are in high demand to enable small molecule library management and processing<sup>[187]</sup>. To this end many, cheminformatics tool kits and scripting frameworks<sup>[188,189]</sup> have been developed such as chemf<sup>[190]</sup>, RDKit<sup>[191]</sup>, CDK<sup>[192]</sup>, and OpenBabel<sup>[193]</sup>.

#### 1.2.1 Small Molecule Library Management

The management of a small molecule library comprises conversion, canonicalization, and normalization of molecular structures as well as the application of search and descriptive algorithms to filter and characterize small molecule libraries<sup>[194]</sup>. Functionality includes, *inter alia*, the removal of mixtures, inorganics, and salts, tautomer normalization, pH calculations, substructure searches, and descriptor calculations.

Cheminformatics libraries, such as the afore mentioned Chemistry Development Toolkit (CDK), offer functionality for the bulk of cheminformatics tasks and are

## 1. GENERAL INTRODUCTION

---

consumed by front-end tools for user interaction<sup>[195]</sup>. In addition, specialised services exist that focus on single steps such as parsing IUPAC names (OPSIN<sup>[196]</sup>) or cross-reference and identifier tracking (UniChem<sup>[197]</sup>). Because the library management process involves many different steps and software, expert tools have become popular that aggregate different software to facilitate the process, hence reducing the number of tools and steps a cheminformatician has to deal with<sup>[198,199]</sup>.

### 1.2.2 Representation of Small Molecules

Representation concerns the storage, transfer, and visualization ability of small molecular structures. Over the last 60 years, various systems have been proposed, of which some have become accepted community standards<sup>[200]</sup>. Representations should encode all relevant information about a chemical structure while being as concise as possible and while maintaining efficient readability for either humans or computers. The choice of representation affects speed, resource requirements, and data handling of cheminformatics tools.

#### Notations and Conventions

File formats represent small molecular structures in a precisely defined way and serve as the smallest unit for structure storage and transfer. The most fundamental chemical file formats are described.

The Simplified Molecular Input Line Entry System (SMILES) is one of the most commonly used line notations. Older notations include the less prominent Wiswesser Line Notation (WLN)<sup>[201]</sup> and the DARC system<sup>[202]</sup>. SMILES encodes a molecular structure in a single sequential character string that is both human- and machine-readable. In contrast to other notations mentioned herein, no comprehensive and formal specification of the line notation has ever been published. For this reason, different implementations of SMILES can differ in functionality, making SMILES unreliable if used across different cheminformatics toolkits. This – taking into account that no official canonicalisation model exists

## 1. GENERAL INTRODUCTION

---

either – is the biggest weakness of SMILES. Issues around SMILES are being addressed by developments in the community sector (OpenSMILES as part of the Blue Obelisk group<sup>[203]</sup>) and academia<sup>[204]</sup>.

A MDL molfile contains a redundant connection table that stores atom and bond connectivity in an atom and bond block respectively. The blocks contain all relevant information about the structure such as charge, atom stereo parity, and valence. Structure-Data files (SDfile) extend MDL molfiles, accommodating any number of molecules in a single file.

InChI, the IUPAC International Chemical Identifier, is a standardized open source line notation that uses layers to represent different levels of chemical structure information<sup>[205]</sup>. These layers encompass constitution (atoms and bonds), charge, stereochemistry, isotopes, fixed hydrogens, and reconnections, i.e. reconnected atoms such as coordinated metal atoms.

The line notation comes in two flavours: the InChI itself and a 27 character-long hashed representation called InChIKey – a more condensed representation of the full InChI targeted at database queries. The standardized IUPAC InChI guarantees proper interoperability across different platforms. Its application is currently limited by its range of unsupported structures, e.g. polymers, Markush structures, mixtures, conformers, and topological isomers<sup>[206]</sup>.

The Chemical Markup Language (CML) is a XML-based file format that encodes a chemical structure in a highly human readable way<sup>[207]</sup>. Extending XML, CML is customisable to accommodate any additional information about the structure in a pre-defined manner. This makes CML useful for problems encountered in the area of data persistence.

### Graphical Representation

Graphical representations of small molecules act as interface between the digital internal representation of a molecular structure and the user. Therefore the graphical representation needs to depict the molecule in its entirety and correctly. This statement is true in most cases for the depiction of molecular graphs with

## 1. GENERAL INTRODUCTION

---

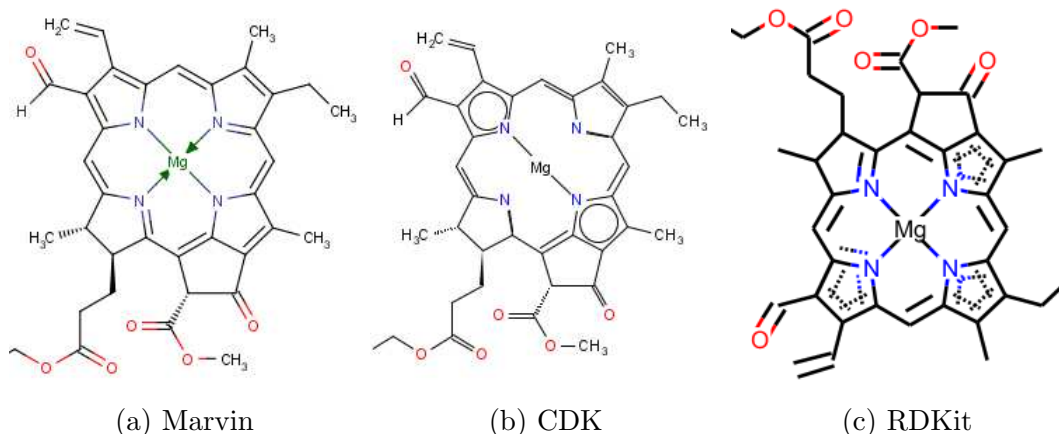


Figure 1.5: Comparison of graphical representations of chlorophyll *f* [CHEBI:61290] by (a) Marvin, (b) KNIME-CDK, and (c) RDKit. The aliphatic ester chain is not shown for depiction purposes. The coordination bond shown in Marvin is ignored in KNIME-CDK and shown as covalent bond in RDKit. Aromaticity is not visually indicated by Marvin but, partially, in KNIME-CDK and RDKit using circles and dashed bonds respectively.

regard to connectivity. In the case of coordination bonds, depictions already start to vary from one toolkit to another. More subtle problems occur with the depiction of stereochemistry and aromaticity, where representations can be misleading for the unprepared user (Figure 1.5).

Whereas most graphical representations of organic molecule are correct and efforts focus on visualisation of molecule clouds<sup>[208]</sup> or chemical space<sup>[209,210]</sup>, it is important to be aware of current limitations to efficiently deal with single small molecules in cheminformatics.

### 1.2.3 Properties of Small Molecules

Descriptors are used to study physicochemical properties of small molecules. Applied in modelling, those quantitative structure-property relationships (QSPR) enable the clustering and prediction of molecular properties<sup>[194]</sup>. Descriptors can be grouped into four classes, where:

- topological descriptors describe properties of the molecular graph in 2D

## 1. GENERAL INTRODUCTION

---

- geometrical descriptors describe properties of the molecular structure in space (3D)
- electronic descriptors describe the energy and charge state of molecular structures
- hybrid descriptors are combinations of the other classes of descriptors

Hundreds of descriptors exist in different toolkits that are often used in combination for model building<sup>[211]</sup>. Subtle differences in implementations of the algorithms and differing molecular representations give QSPR descriptors an involuntary cross toolkit complementarity. Because descriptors act on the *internal* molecular representation of a molecular structure to describe its physicochemical properties, any cheminformatics toolkit needs to exercise great care when it comes to structure conversion into its own molecular representation in order to configure the molecular structure correctly for QSPR descriptor calculations. For example, utilization of the same aromaticity model across all molecules in a library is essential for consistent results.

### 1.2.4 Workflow Environments for Cheminformatics

Workflow environments have become increasingly popular over the last decade with the promise to simplify integration and coordination of different software packages<sup>[212,213]</sup>. Researchers often face the challenge of processing and analysing complex data sets. This involves the use of various tools, frequent saving and loading of data in different formats, and data transformation or manipulation<sup>[131]</sup>. In addition, these activities should be recorded to ensure reproducibility and extendibility. Workflow tools address the problem of orchestrating these processes and offer a potential all-in-one solution, bringing together numerical, textual, chemical, and biological data<sup>[214]</sup>. The concept behind those tools can be understood as *visual programming*<sup>[215]</sup>.

This introduction concentrates on platforms that support bioinformatics and cheminformatics tasks, not on generic workflow environments for business intelligence. The first group comprises Galaxy<sup>[216]</sup>, Taverna<sup>[217]</sup>, KNIME<sup>[218]</sup>, and

## 1. GENERAL INTRODUCTION

Pipeline Pilot<sup>[219]</sup>. Galaxy is a web-based platform that focuses on genomic data and offers only rudimentary cheminformatics functionality, e.g. in the form of “ballaxy”<sup>[220]</sup>. Taverna, KNIME, and Pipeline Pilot are desktop applications, with the latter being the *de facto* standard for cheminformatics. Developed by Accelrys, Pipeline Pilot has been specifically developed for bio- and cheminformatics needs in life sciences. In contrast, the free-of-charge open source workflow management systems Taverna and KNIME are more generically targeted at workflow generation for data transformation, largely relying on contributions from the scientific community in the form of plug-ins and shared workflows<sup>[221]</sup>.

Most workflow platforms follow the same principle (Figure 1.6). Tasks or processes are carried out by discrete entities, that – based on the platform – are

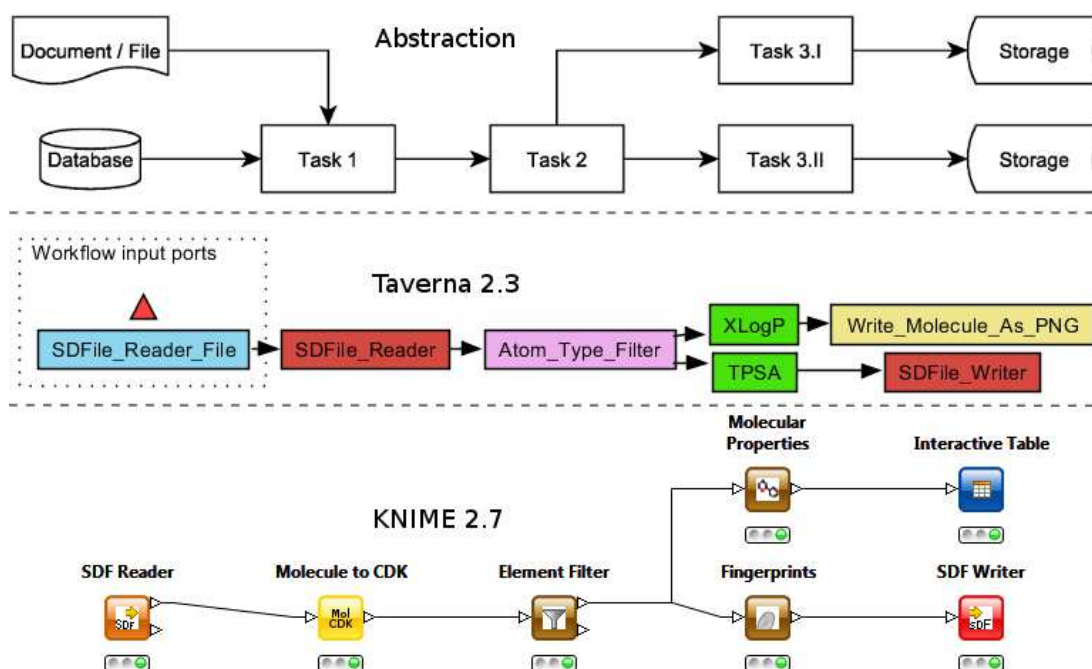


Figure 1.6: Schematic of the principle mechanism of workflow platforms. Data is either loaded from external sources or created within the workflow environment. Loaded data is then sequentially passed on to individual entities that carry out their tasks (Task 1, Task 2, *et cetera*). Workflows can include branches as indicated with Task 3.I and Task 3.II, where different intermediate results are generated. Task results are stored for persistence or usage outside the workflow environment. The middle and lower part of the figure depicts workflows that follow the same pattern from Taverna v2.3 and KNIME v2.7 respectively.

## 1. GENERAL INTRODUCTION

---

called nodes, workers, components, *et cetera*. These entities either create input, e.g. by reading a file or querying a database, or take input from another entity. It follows that entities can also output data – typically after an operation has been applied on the data – or remove data by storing it outside the platform. The way in which data is transferred from one entity to another depends on the platform. For example, the transfer could either be file based or tabular. The parameters of a specific function are typically set through a configuration dialogue of an entity and define the behaviour of the function for that entity. A workflow is made up of individual entities that are connected to each other under the constraints of their input and output requirements. Workflows are intrinsically linear, i.e. execution flows from an input to an output operation, but allow for branching. More advanced structures, such as loops, are supported in some environments, rendering the pipeline design process more flexible. On workflow execution, entities in a pipeline execute either sequentially – one after another – or in batch, where pieces of data are processed downstream when they become available. Intermediate results can be stored for inspection.

In summary, workflow environments enable scientists to build complex data processing and analysis pipelines, record how the data is processed, and share their workflows, all through the concept of visual programming. Whereas properties like user-friendliness, ease of debugging, and the ability to inspect intermediate results are distinct advantages of workflow platforms, disadvantages come from the limited functionality offered, forcing scientists to go beyond their pipelines, the need to cache data impeding scalability, and the initial overhead of implementing new entities<sup>[214]</sup>.

### **Konstanz Information Miner**

The Konstanz Information Miner (KNIME)<sup>[218]</sup> is a Java-based open-source workflow platform that supports a wide range of functionality. It has an active developer community with plug-ins for bio- and cheminformatics<sup>[150,222–224]</sup>. For a detailed description of the KNIME data analysis platform see<sup>[225]</sup>.



## 1. GENERAL INTRODUCTION

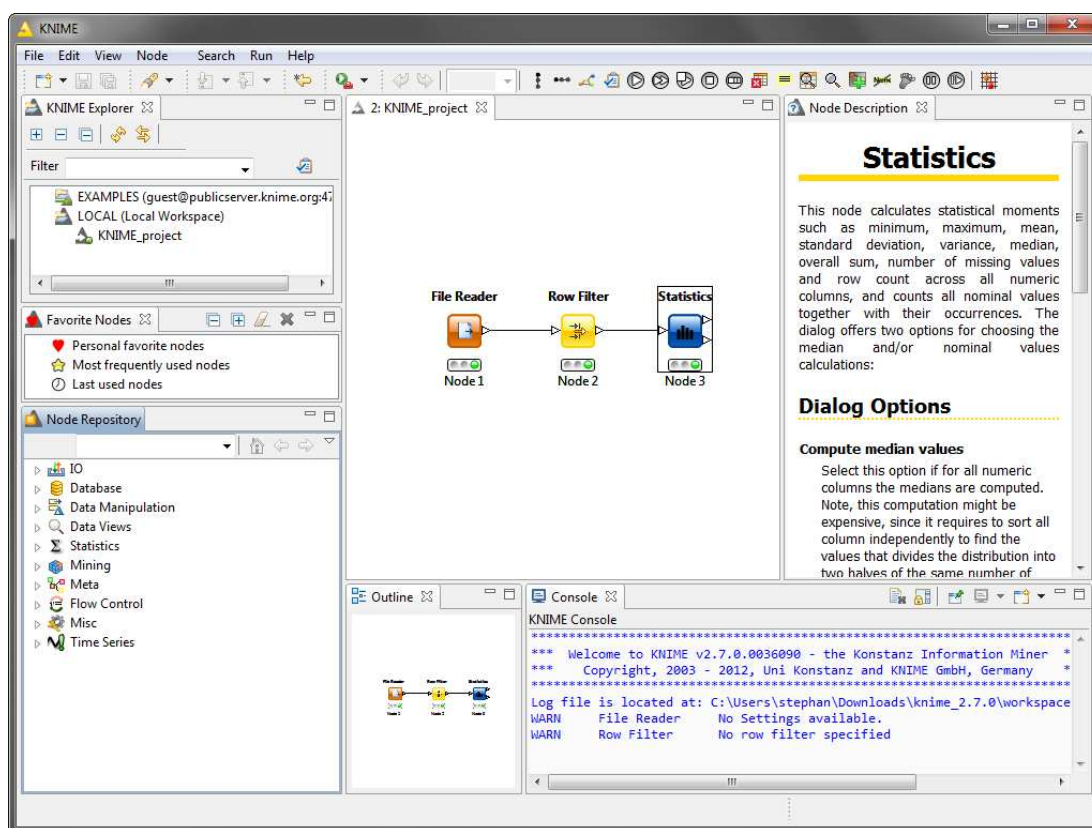


Figure 1.7: Screenshot of the KNIME workbench with an opened project. The Editor in the middle serves as work pane where workflows can be build via addition of nodes from the Node Repository via drag-and-drop. Supporting information about a selected node is displayed in the Node Description pane. The KNIME Explorer and Console hold the list of workflows and system message respectively.

The KNIME workbench follows a simple layout, as can be seen in Figure 1.7, comprising the following components: A KNIME Explorer holding all workflows, a Node Repository listing all available entities (nodes), a Console that displays system messages, a Node Description that displays additional information about any selected node, and an Editor where individual nodes can be put together to build a workflow. Nodes can be added to the editor pane via drag-and-drop from the Node Repository.

KNIME follows a sequential execution pattern. Nodes in a pipeline are executed one after another after previous nodes have finished execution. Intermediate results are cached and, depending on their size, serialized to disk. Nodes exchange

## 1. GENERAL INTRODUCTION

---

data in tabular format with rows and column, where every column must have a data type specification. Rows consist of one to many cells that span one to many columns. A cell is an atomic unit of information that cannot be split further. Each cell has a data type that defines its content based on the column it is in. It is important to note that KNIME nodes can only deal with cell types that match their input requirements.

Internal KNIME plug-ins are available via the standard in-built repository. External third party plug-ins can either be added from an archive file or update site. An official community repository and forum is available to support the development of third party plug-ins.

KNIME is developed in Java version 1.7 and is based on the Eclipse framework, an integrated development environment. Consequently, individual nodes and full plug-ins can be developed against the Eclipse framework and the Java Software Development Kit. The development process is guided by a node extension wizard that creates four Java classes (Table 1.2). These four Java classes determine the node behaviour and are sufficient for node development. More advanced structures are documented in the KNIME Application Programming Interface.

Java class	Responsibilities
NodeModel.java	validates dialog and node port parameters, saves/loads internal settings, and defines the main function body
NodeDialog.java	defines the settings window and node parameters
NodeView.java	defines a custom view of the stored data
NodeFactory.java	orchestrates instantiation of the model, dialog, and view
Node.xml	markup-language based node description

Table 1.2: Java classes for KNIME node development. The KNIME node extension wizard facilitates the development of external nodes. The four base classes created by the wizard are listed including their main responsibilities, i.e. the behaviour that can be defined. In addition, a markup language file is created that contains meta information about the node.

### 1.3 Aim of this Thesis

The aim of this thesis is to provide effective yet usable software tools that simplify data analysis and exploration for the general scientific community. By narrowing the gap between metabolomics data and metabolomics data analysis, I attempt to bridge life science and computer science. Informed data processing and analysis is a difficult problem that demands expertise knowledge and time-intense manual labour. The plethora of data available and the accelerating pace at which new data is acquired, increase the demand for semi- or fully-automated pipelines that make the data analysis process more manageable and efficient.

The non-linear, multi-parametric metabolic responses captured in metabolomic snapshots – typically taken from two or more sample groups – contain crucial information. Information that could be used in early-stage detection of disease. However, the quality of the available data depends on the design of a study, its sample set-up, instrumental configuration, *et cetera*. These factors demand attention to detail from an analyst, but also require adequate software to offer support<sup>[3]</sup>.

To achieve the aim of narrowing the gap between metabolomics data and metabolomics data analysis, the objectives are to develop a modular library for processing of LC-MS<sup>n</sup> data and integrate this library as external plug-in in a workflow environment. Methods for LC-MS<sup>n</sup>-based metabolite identification are to be implemented to broaden the scope of the application and address the need for identification frameworks that go beyond current metabolomics databases<sup>[17,38]</sup>. *Visual programming* is to be explored as a method of choice to make these data processing functions accessible to analysts without programming experience and ensure reproducibility of analysis<sup>[226]</sup>: the building-block approach of the chosen workflow platform, KNIME, is to serve as platform to combine the tool's functionality with more generic, already existing, methods. In addition, cheminformatics methods are to be explored to enable small molecule management, ultimately feeding back into LC-MS<sup>n</sup> analysis.



# INFORMATICS FOR LC-MS<sup>N</sup> ANALYSIS

---

## 2.1 Introduction

Metabolomics studies aim to characterise biological samples and identify metabolites<sup>[11,12]</sup>. To investigate the small molecule complement of organisms, liquid chromatography coupled to tandem mass spectrometry (LC-MS<sup>n</sup>) is a routine technique commonly used. LC-MS<sup>n</sup> is applied in profiling, fingerprinting, or untargeted mode<sup>[14]</sup> in a variety of areas including environmental<sup>[39]</sup>, plant<sup>[58,63]</sup>, and biomedical research<sup>[227]</sup>. Modern LC-MS<sup>n</sup> systems can detect more mass traces than ever before thanks to high mass accuracy ( $m\Delta \leq 2$  ppm<sup>[33]</sup>) and resolution ( $R \geq 100,000$ <sup>[34]</sup>), producing complex, information-rich data for every sample.

Data typically consists of mass-to-charge ( $m/z$ ), time, and intensity triplets that describe for every detected ion mass the strength of the ion beam and the time it is detected by the spectrometer. Processing and interpreting these data matrices is extraordinarily difficult because of the high dynamic range, chemical diversity, and metabolite numbers typically found in metabolome samples. The partially

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

convoluted, densely populated signal landscape contains systematic and random noise amongst true signals of varying intensity and shape. Additionally, formation of ion clusters, adducts, or fragmentation implies that many of the extracted peaks or features can belong to the same compound.

Both proprietary and free libraries addressing the data processing problem have been developed and are routinely applied in LC-MS<sup>n</sup> metabolomics<sup>[37,175]</sup>. Due to the inherent nature of metabolomics data, processing and analysis requires complex workflows, bundling different programs, traversing parameter space, pulling in additional information from databases, and performing statistical multivariate analysis. Consequently, pipelines have been built concatenating existing tools<sup>[93,155,171,228]</sup>.

Workflow platforms such as the Konstanz Information Miner (KNIME)<sup>[218]</sup> offer the potential for an all-in-one solution. OpenMS<sup>[150]</sup>, a library for LC-MS data management and analyses, primarily geared towards proteomics, has already been added to KNIME's bioinformatics suite. Workflow-based data processing can be described as visual programming. It has the advantage of ease-of-use for computational and experimental scientists alike and enables rapid development of complex pipelines while maintaining flexibility due to modularity.

We have developed MassCascade and its plug-in MassCascade-KNIME, a library and node-suite for stepwise LC-MS<sup>n</sup> metabolomics data processing. In the following sections, we give an overview of the architecture of the library and plug-in, summarise the implemented features, and demonstrate the performance and advantages of a unified workflow environment for fingerprinting.

### 2.2 MassCascade's Implementation

The library MassCascade comprises various methods for data processing, visualisation, and feature identification. Each method is implemented using Java's<sup>®</sup> concurrency framework<sup>[229]</sup> for multi-threading to increase execution speed. The library is thread-safe and can be executed in a server environment. MassCascade has been developed in the programming language Java<sup>®</sup> version 1.7.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

### Structure

MassCascade works based on a set of essential instances that represent abstract mass spectrometry entities. These core instances are passed between processing methods (Figure 2.1). The *MS Data* type contains raw data taken from individual scans, the *Feature Data* type contains extracted features and associated annotations, and the *Feature Set* type contains a collection of features that are correlated, i.e. compound spectra. They are defined as follows:  $S$  (scan) is a set of  $m/z$ -intensity value pairs  $s_i$  at a given acquisition time  $t$ :

$$S_t = \{s_1, s_2, s_3, \dots, s_m\} \quad (2.1a)$$

$$s_i = (m/z, I) \quad (2.1b)$$

For a given  $m/z$ ,  $F$  (feature) is a triplet containing a time vector  $\vec{t}$ , an intensity vector  $\vec{I}$ , and a retention time  $rt$ . The time and intensity vector are of identical length and represent the chromatographic profile of the feature. The retention time is the characteristic time of the feature, typically indicating the apex of the chromatographic profile.  $FS$  (feature set) is a set of features  $F$  at a time point  $t$ , which is the consensus ‘retention time’ of the complete feature set:

$$F_{m/z} = (\vec{t}, \vec{I}, rt) \quad (2.2)$$

$$FS_t = \{F_{m/z_1}, F_{m/z_2}, F_{m/z_3}, \dots, F_{m/z_n}\} \quad (2.3)$$

Each method takes a set of parameters including one or many MS instances, applies the method’s function on the instance, and returns a new MS instance. It is apparent that those instances essentially are data containers. This way a snapshot of the data can be serialised to disk after any processing step if required. That is essential for workflow environments, where intermediate results need to be accessible after every execution step.

Too many disk input and output operations are not desirable however because they slow down execution. Unless constrained by computer memory, the library can be run in memory mode. All MS instances can either be file- or memory-based, i.e. their underlying data is serialized to external files or kept in memory

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

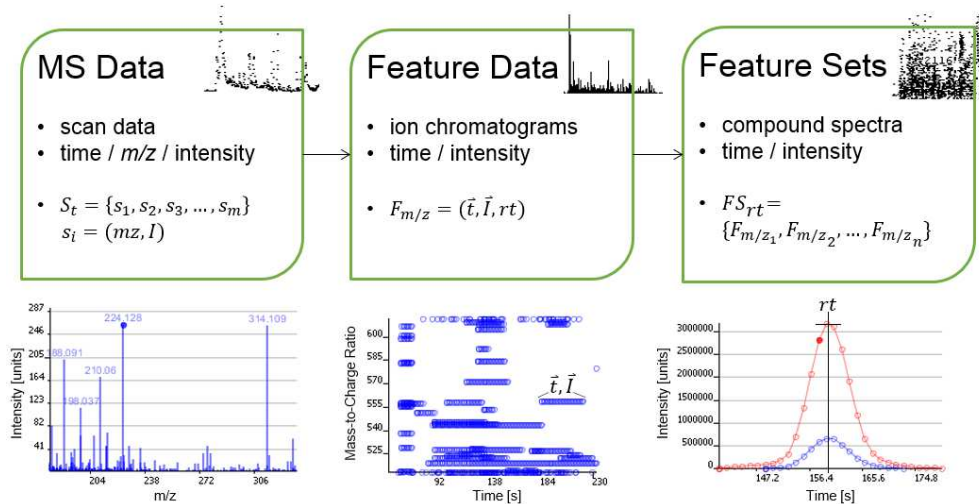


Figure 2.1: Overview of MassCascade data types and their representations. The *MS Data* type contains raw data taken from individual scans:  $S$  (scan) is a set of  $m/z$ -intensity value pairs  $s$  at a given scan time  $t$ . The *Feature Data* type contains extracted features and associated annotations: for a given  $m/z$ ,  $F$  (feature) is a triplet containing a time vector  $\vec{t}$ , an intensity vector  $\vec{I}$ , and a retention time  $rt$ . The *Feature Set* type contains feature sets  $FS$ : a set of features  $F$  at a time point  $t$ .

(Figure 2.2). The latter mode is desirable for server-side applications because of its increased execution speed. To facilitate file- or memory-mode selection, respective `ContainerBuilder` have been implemented that automatically propagate the correct mode from one method to another when a new container is generated. The builders only need to be explicitly utilized once at the beginning of the execution script.

### Core Framework

The core library contains methods for LC-MS<sup>n</sup> data processing. As outlined above, every method takes an instance of a data type and returns another instance of the same or a different data type. For example, the `ProfileBuilder` method accepts a *MS Data* instance, extracts features, and returns a *Feature Data* instance. Each method is defined as predefined constant (*enum type*) in `CoreTasks` that defines the method's Java<sup>®</sup> class and unique identifier. By default, each method also extends `CallableTask`. This interface represents the



## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

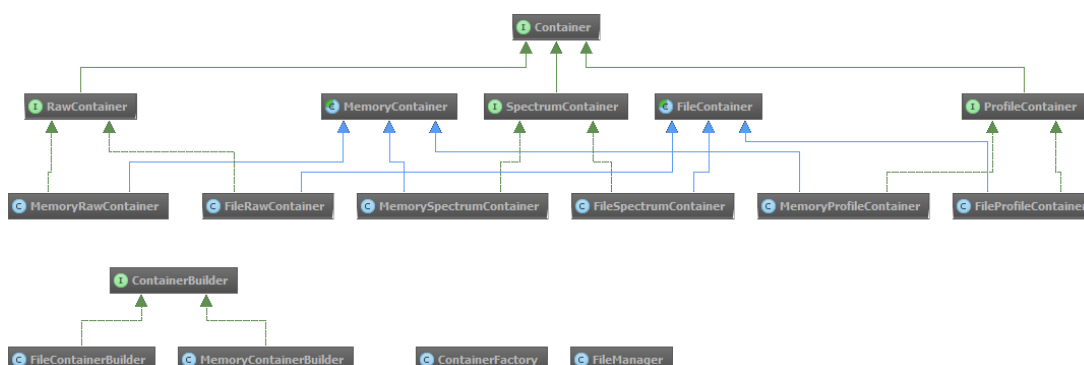


Figure 2.2: UML diagram of MassCascade’s data structure. The library can run in either file- or memory-mode, determining whether the data of the employed data containers is serialized to disk or kept in memory. For every data type, a file- and memory-container exists. The respective `FileContainerBuilder` and `MemoryContainerBuilder` ensure automatic propagation of the selected mode when different methods of the library are used in succession.

contract that all methods must follow within the core framework. This design pattern guarantees seamless usage and integration of the tool in various environments.

Input and output parameters are clearly defined in the methods’ annotations. Annotations are provided in the form of JavaDoc for the case of the core library. Parameters can be passed to methods *via* a `ParameterMap` as `Parameter`-value pairs (Figure 2.3). This highly verbose way of adding parameters to methods facilitates easy usage and readability of written code, essential for maintainability. Parameters are centrally defined in the `Parameter` class with a short description and – where applicable – default values.

### Visualisation Framework

The visualisation framework enables data inspection for data type instances. It comprises a spectrum viewer to visualise spectra and chromatograms. Textual annotations can be added on demand. Data traces can be displayed with various styles: impulse, line, polyline, spline, and point. The viewer also offers common interactive viewing functions: zooming, highlighting, and coordinate labels.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

```
public class PsiMzmlReader
extends CallableTask

Class for reading PSI mzML files.
• Parameter DATA_FILE - The data file to be read.
• Parameter RAW_FILE - The target raw container.

// build the output data container
RawContainer container = MemoryContainerBuilder.getInstance()
    .newInstance(RawContainer.class, file.getName());

// build the parameter map
ParameterMap params = new ParameterMap();
params.put(Parameter.DATA_FILE, file);
params.put(Parameter.RAW_CONTAINER, container);

// create and run the task
CallableTask task = new PsiMzmlReader(params);
RawContainer outContainer = (RawContainer) task.call();
```

Figure 2.3: Example of JavaDoc documentation for the `PsiMzmlReader` method (left) and minimalistic source code to read a data file in mzML file format (right). The parameter requirements of the reader method are defined in its documentation. A memory-based instance is chosen to serve as container for the read raw file. The file and data type instance are passed to the reader method *via* a `ParameterMap`. The map uses pre-defined parameters that correspond to the parameters defined in the documentation.

In addition, the visualisation framework contains a set of pre-defined tables for block-like use in software applications that consume the library. Caching and lazy loading are have been used to enable responsive user-table interactions, e.g. with large numbers of features loaded.

### Identification Framework

The identification framework contains methods for metabolite identification, including web-based and local services for metabolite and spectra databases. Extracted features can either be queried *via*  $m/z$  lookups or spectra queries. Pseudo spectra contained in the *Feature Set* type serve as spectra for the latter type.

After databases lookup, putative metabolite identifications can be ranked based on isotope, adduct, and tandem mass spectrometry using a custom scoring scheme. Aggregated information is used to filter out irrelevant metabolites and narrow down the search space. Identified metabolites and signal lists can be exported for further analysis with other programs.

### 2.3 MassCascade’s Functionality

MassCascade’s methods are grouped into pre-processing, processing, and post-processing. Pre-processing deals with the manipulation of raw data up to the point of feature extraction. Processing addresses the extracted features and ends with post-processing, the annotation and identification of signals or compound spectra. Throughout development, care was taken to incorporate adaptive approaches that work close to the raw data and avoid binned data. Binned data offers many advantages including, *inter alia*, increased execution speed and lower memory costs, but introduces the danger of lowering the granularity of high-resolution data beyond informative thresholds, thus deleting information. Whenever binned data is a prerequisite algorithmically, equidistant binning is utilized with non-flexible boundaries. Equidistant binning creates equally spaced bins. Values that fall within a particular bin are integrated and assigned to that bin. Default parameters are provided for all implemented methods but these may need refinement based on the quality of the data.

#### 2.3.1 Data Pre-Processing

The library supports the HUPO PSI mzML 1.1.0 specification<sup>[96]</sup>, superseding the older mzXML<sup>[94]</sup> and mzData<sup>[95]</sup> formats, and Thermo Scientific’s RAW file format. The methods have been optimised for centroid data. Input data should be centroided with one of the many available file converters such as ProteoWizard<sup>[230]</sup> or by using the implemented wavelet-based centroider after profile data has been read in. The wavelet method uses a Ricker wavelet of the form:

$$\psi(t) = \frac{2}{\sqrt{3\sigma\pi^{\frac{1}{4}}}} \left(1 - \frac{t^2}{\sigma^2}\right) e^{\frac{-t^2}{2\sigma^2}} \quad (2.4)$$

$\psi$  denotes the wavelet with offset  $t$  and spread  $\sigma$ . The Ricker wavelet is the negative normalized second derivative of a Gaussian function. This wavelet is convoluted over every scan to find centroids in the profile signals, assuming the ideal case where these follow a Gaussian distribution<sup>[231]</sup>.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

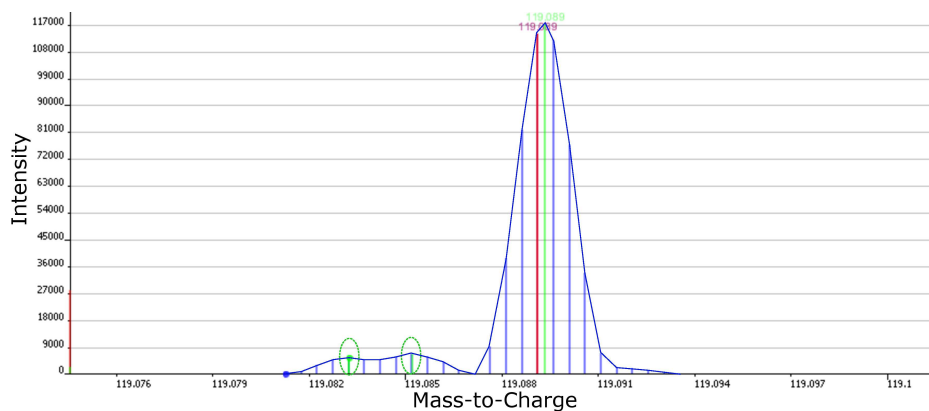


Figure 2.4: Overlay of a section of a scan in profile mode (blue, with envelope), wavelet centroided (red), and vendor centroided (green). The vendor methods picks up two additional centroids (highlighted) and – for the largest centroid – sits at the apex of the profile envelope. The wavelet algorithm finds only one centroid and chooses the closest discrete signal of the profile set, resulting in a  $m/z$  deviation of 0.0002.

The resulting centroid of a single profile does not represent an interpolated centroid from the apex of the distribution but the closest discrete signal taken from the profile (Figure 2.4). This can result in a small  $m/z$  error in the region of  $\Delta m/z \approx 0.0001$ .

Systematic and random noise reduction are supported following an approach outlined by Zhu *et al.*<sup>[102]</sup>. The algorithm consists of two parts:

1. Systematic noise (background) can be subtracted through a blank reference sample containing only solvent. Instrument-dependent drifts and background ion traces, e.g. from the solvent or ubiquitous contaminants, are represented in these blank injections. Subtraction of a blank reference from a sample effectively removes these spurious traces. This method assumes that systematic noise or background is linearly added to the desired signals and can thus be removed *via* subtraction of a blank reference. For every data point  $s_i$  in every scan of a sample, a two-dimensional control area is created in the blank reference for a given time range  $\Delta t$  and  $m/z$  tolerance range  $\Delta m/z$  (Figure 2.5a). A data point is considered noise and removed

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

if a similar data point  $s_j$  is present in the control area of the reference:

$$(m/z)_{s_i} - \Delta m/z \leq (m/z)_{s_j} < (m/z)_{s_i} + \Delta m/z \quad (2.5)$$

$$t_{s_i} - \Delta t \leq t_{s_j} < t_{s_i} + \Delta t \quad (2.6)$$

2. Data points resulting from random noise are removed by inspecting adjacent scans in the same sample for similar data points (Figure 2.5b). A data point  $s_i$  is considered noise and removed if it does not have a set of adjacent neighbours  $S = s_j, s_{j+1}, \dots, s_{j+n}$  within a given  $m/z$  tolerance  $\Delta m/z$  and the total length of the chain of adjacent neighbours, i.e. the length of the putative ion trace, falls below a scan number threshold  $n$ . In addition, the data point or at least one of its adjacent neighbours has to exceed a minimum intensity threshold  $I_{min}$ :

$$(m/z)_{s_i} - \Delta m/z \leq (m/z)_{s_j} < (m/z)_{s_i} + \Delta m/z \quad (2.7)$$

$$length(S) \geq n \quad (2.8)$$

$$I_{max}(S) \geq I_{min} \quad (2.9)$$

Data filters operating in all three domains – time,  $m/z$ , and intensity – allow raw data trimming. The chromatographic element of a LC-MS system can be significantly inconsistent at the beginning and end of each run due to column properties and the mobile phase mixture typically used at those phases. Regions above or below a certain  $m/z$  value may not be of interest experimentally or signals below a certain minimum intensity that have not been filtered out by the instrument may be considered noise *a priori*. Data filters crop the data based on threshold values or ranges and remove unwanted or unreliable data segments.

Feature extraction refers to the detection and isolation of distinct ion traces. Features are build using a simple adaptive incremental approach. Scans are visited in chronological sequence. Every data point in a scan is either assigned to a previously created ion trace – originating at an earlier scan – or becomes the source of a new ion trace. Only one data point per scan be assigned to an ion trace.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

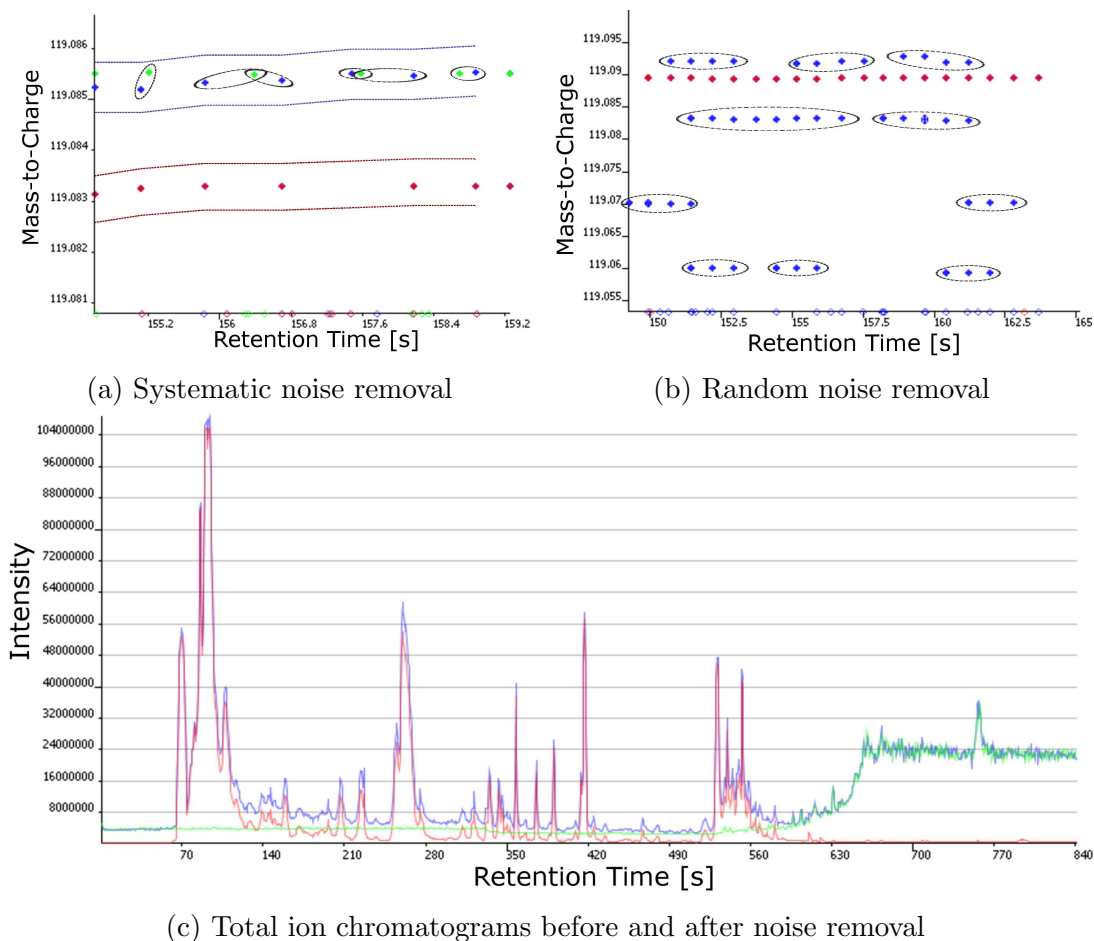


Figure 2.5: Illustration of noise removal pre-processing methods. (a) Time versus  $m/z$  plot of signals pre (blue) and post (red, covering blue) systematic noise removal. Blue and green bands indicate the  $m/z$  tolerance window of the blue and green trace. An ion trace of the background reference (green) falls within the window of the upper trace around 119.085. Consequently, the sample ion trace is removed. (b) Time versus  $m/z$  plot of signals pre (blue) and post (red, covering blue) random noise removal. Traces lower than eight successive scans are removed (circled). (c) Total ion chromatograms of the pre- (blue), post-processed (red) sample and background reference (green). Noise reduction dramatically reduces baseline distortions while leaving large features intact.

Assignment is conditional on the  $m/z$  value of the data point being closest to the weighted average of all  $m/z$  values captured by the ion trace from previous scans and falling within the given  $m/z$  tolerance calculated around the trace's weighted average (Figure 2.6). The weighted average is updated on every assignment based

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

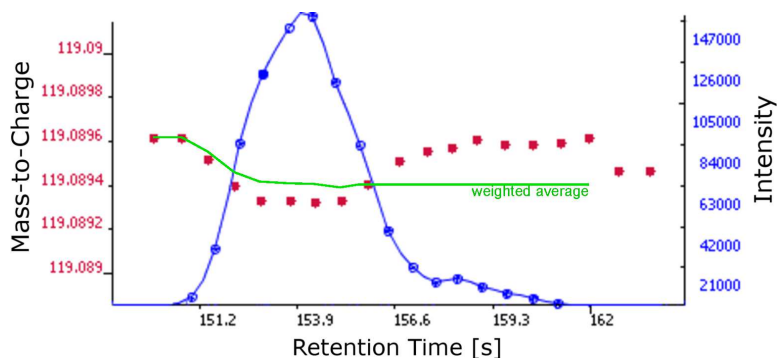


Figure 2.6: Illustration of the feature extraction process. Data points of the ion trace  $m/z = 119.0984$  are shown with regard to their individual  $m/z$  values (red) and intensity (blue). The red points demonstrate the  $m/z$  measurement error. The course of the characteristic  $m/z$  value determined by a weighted average is shown in green. After the trace's maximum, the value is set to 119.0984. Subsequent low-intensity values only introduce marginal deviations.

on the data point's intensity, placing higher emphasis on high-intensity, i.e. more accurate,  $m/z$  values. Finally, an ion trace qualifies as feature if it exceeds a given minimum length (number of adjacent data points) and the distribution's apex is above a minimum intensity threshold.

To eliminate baseline drifts in the absence of a blank reference, a morphological Top Hat filter has been implemented acting on individual features. The specific implementation is described by Herk *et al.*<sup>[232]</sup>. Top Hat by opening is applied, i.e. erosion followed by dilation, with a structuring element  $B$  on a feature  $f$  (Figure 2.7). Erosion shaves off the peaks and reduces their width. Following, dilation widens the flattened peaks. It reconstructs the shape without the shaved off peaks. By applying dilation after erosion, i.e. opening, the baseline under the peaks is first estimated and can subsequently be subtracted from the original feature. The structuring element is characterised by a window width in the time domain where the baseline drift occurs. The element itself can be understood as a simple window that defines the granularity of the method. A wider window width is more coarse. Peaks within the window width are flattened and the resulting baseline estimate is more conservative.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

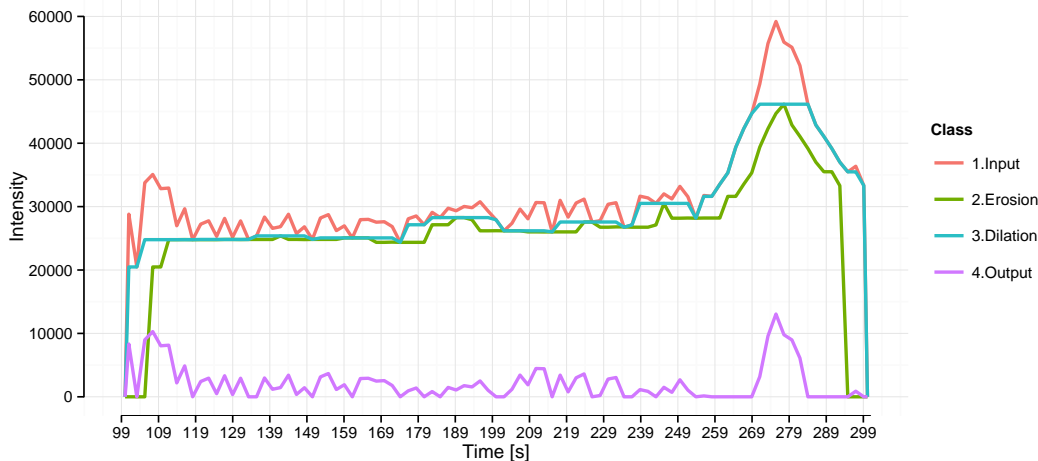


Figure 2.7: Illustration of the TopHat algorithm. The input chromatogram (1.Input) is eroded with a structuring element of length six resulting in the ‘peak-shaved’ green line (2.Erosion). Dilation takes the eroded data and reconstructs the original shapes without the peaks (3.Dilation). Subtraction of the estimated baseline (3.Dilation) from the original input (1.Input) yields the baseline-corrected chromatogram (4.Output).

$$\text{Dilation : } (f \oplus B)(x) = \max \{f(x - x') | x' \in B\} \quad (2.10)$$

$$\text{Erosion : } (f \ominus B)(x) = \min \{f(x + x') | x' \in B\} \quad (2.11)$$

$$\text{Opening : } f \circ B = (f \ominus B) \oplus B \quad (2.12)$$

$$\text{TopHat : } f' = f - (f \circ B) \quad (2.13)$$

Note that  $\oplus$  and  $\ominus$  indicate operations that should be understood as additions and subtractions in a non-numeric context. Here, the application of the structuring element widens and reduces the peak shapes. The symbol ‘ $\circ$ ’ indicates function composition. The baseline subtracted from the original feature gives the new baseline-corrected feature. The remaining features, after noise reduction, feature extraction, and – if required – baseline correction, still contain many biologically irrelevant features. To further narrow down the solution space to extract only high-quality ion traces, two criteria have been implemented: The *mass chromatographic quality* from the Component Detection Algorithm (CODA) and the *Durbin Watson score* based on a combination of CODA with a Durbin Watson statistic<sup>[233]</sup>.



## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

The mass chromatographic quality ‘*mcq*’ describes the similarity between a raw (length-scaled  $x(\lambda)_i$ ) and a smoothed (auto-scaled  $x(r, s)_i$ ) version of a feature.  $x_i$  denotes the intensity value at position  $i$  in the feature’s intensity vector.  $\lambda$ ,  $s$ , and  $r$  indicate that the value is length-scaled or standardised and smoothed. Higher quality features that are less likely to be random variations have a score closer to one. The smoothed and raw version of the signal region are almost identical. Noisy features show greater discrepancy and give a score closer to zero. The window width  $w$  defines the granularity of the method in the time domain and  $n$  corresponds to the total number of data points of the feature.

$$mcq = \frac{1}{\sqrt{n-w}} \sum_{i=1}^{n-w+1} x(\lambda)_i x(w, s)_i \quad \{mcq \in \mathfrak{R} \mid 0 \leq mcq \leq 1\} \quad (2.14a)$$

$$x(\lambda)_i = \frac{x_i}{\sqrt{\sum_{i=1}^n x_i^2}} \quad (2.14b)$$

$$x(r)_i^R = \frac{1}{w} \sum_{k=i}^{i+w-1} x_k \quad (2.14c)$$

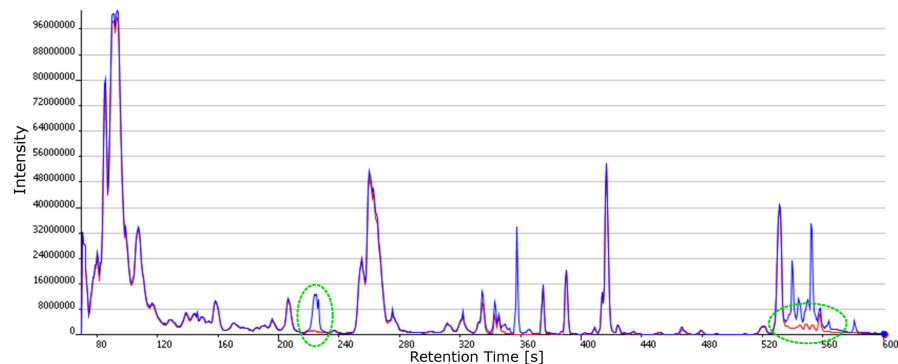
$$x(s)_i = \frac{x_i - \mu}{\sigma} \quad (2.14d)$$

The mass chromatographic quality gives good results for baseline-free ion traces. If solvent signals or baseline artefacts are present, the *mcq* score gets disproportionately worse because the smoothed version of the feature appear to contain more random signal. To reduce this problem, the Durbin Watson score ‘*dw*’ has been implemented as alternative (Figure 2.8). It acts on elements  $x'$  of the first derivative of a feature and summarizes the relative change in the feature through the normalized sum of squared intensities. For noisy features, the relative change will give higher *dw* values. The window width  $w$  defines the granularity of the method in the time domain.

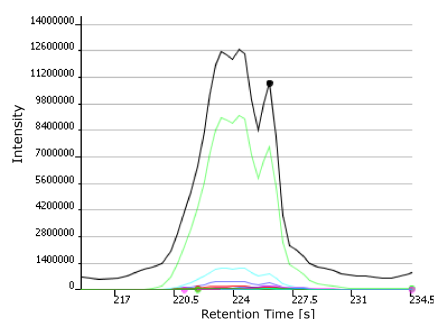
$$dw = \frac{\sum_{i=2}^w (x'_i - x'_{i-1})^2}{\sum_{i=1}^w x_i'^2} \quad \{dw \in \mathfrak{R} \mid 0 \leq dw \leq \infty\} \quad (2.15a)$$

$$x'_i = x_{i+1} - x_i \quad (2.15b)$$

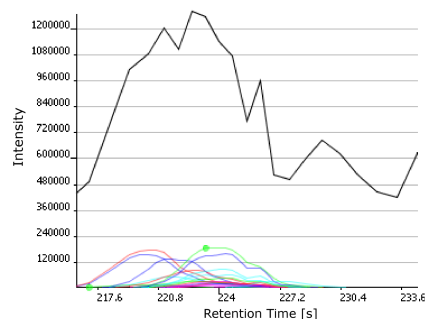
## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS



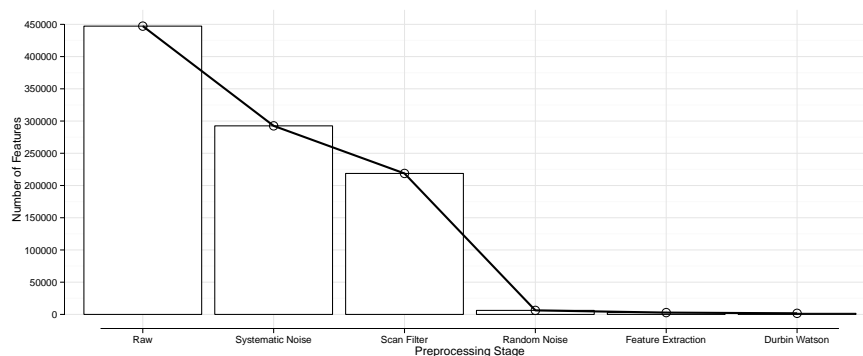
(a) Total ion chromatograms before and after Durbin Watson filtering



(b) (Total) ion traces pre-filter



(c) (Total) ion traces post-filter



(d) Bar chart summarising pre-processing steps

Figure 2.8: Illustration of Durbin Watson filtering and bar chart summary of pre-processing steps. (a) Total ion chromatograms of the pre- (blue) and post-filtered (red) sample using a strict Durbin Watson filter ( $dw = 1.5$ ). The green circles highlight areas of difference. (b) Close up of the unfiltered area around 224 s. (c) Close up of the filtered area around 224 s. In (b) and (c) the total ion trace is shown in black, individual features are shown coloured. Irregular features are removed by the Durbin Watson filter leaving only well behaved features. (d) Bar chart summarising the decrease of the number of features during pre-processing. The 260 fold reduction demonstrates the necessity for pre-processing for efficient data handling.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

The threshold values for the CODA and Durbin Watson filters depend on the individual sample sets and need to be adjusted accordingly. For data exploration, traversal of percentiles of the total score distribution is recommended. This way the impact of lower quality signals, which may be relevant, can be evaluated on the final model.

A Savitzky-Golay filter has been implemented for feature smoothing, increasing the signal-to-noise ratio without distorting the ion trace. For high signal-to-noise data, smoothing can recover ion traces that would be filtered out as noise by other pre-processing methods. The data is assumed to be spaced equally ( $h = 1$ ), allowing the use of convolution coefficient ( $a$ ) lookup tables. A smoothing window of length  $n$  is convoluted over a feature and a polynomial  $Y$  of low degree  $k$  is iteratively fitted through the data points that lie within the smoothing window. The midpoint  $\bar{x}$  of the smoothing window is replaced by the value derived from the fitted function. The window indices are defined by  $z$  where the midpoint  $\bar{x}$  equals zero. Afterwards, intensities of smoothed features are restored to the same magnitude of the raw feature to avoid distortion.

$$Y = a_0 + a_1z + a_2z^2 + a_3z^3 + \dots a_kz^k \quad (2.16a)$$

$$z = \frac{x - \bar{x}}{h} \quad (2.16b)$$

To remove features that are believed to be genuine but not of experimental interest, e.g. contaminants and interferences, features can be filtered by their exact ion mass, further trimming the data set.

### 2.3.2 Data Processing

Data processing addresses extracted features and refines those intra- and inter-sample. Features need to be deconvoluted and peak-picked to isolate compounds of similar or identical mass that eluted at different time points (Figure 2.9a). A popular second derivative polynomial approach and a more robust modified Bieman<sup>[113]</sup> approach have been implemented. For the first approach, a higher order polynomial is fitted through coefficient-lookup. The second derivative of

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

the function determines the boundaries (zero-intercepts) and maximum intensities (local minima/maxima) of individual peaks in the trace. The method is sensitive to the quality of the initial polynomial fit and was found to be less effective for diverse peak shapes. The modified Bieman algorithm for peak picking is purely-shape based: it finds local minima (boundaries) and maxima (peak apices) through iterative data point traversal from the ion trace’s global maximum. The trace is split into individual peaks based on the slope of the declining trace (see Stein *et al.*<sup>[113]</sup>).

For both deconvolution methods, intensity-based thresholds are in place to remove low-level – noisy – signals. For the polynomial approach, the second derivative needs to exceed a pre-defined threshold to qualify as separate peak. the Bieman algorithm carries out a least squares background estimation from the lower half of the total set of background-subtracted ion trace data points.

To remove retention time shifts across samples, Obiwarp – a tool for ordered bijective interpolated warping – can be utilized<sup>[121]</sup>. The parameters required by Obiwarp are extensively documented on its official web page and are mostly self-explanatory, e.g. the parameters for gap initiation and extension penalties. Before alignment of a collection of features of a sample to another collection of feature of a reference, i.e. the adjustment of the feature landscapes, these collections need to be binned in the time and  $m/z$  domain to create a regular grid that is shared by all samples and the reference. Equidistant binning is used for both domains. The smaller the bin size in either domain, the higher the resolution and more fine-grained the resulting alignment. Too small bin sizes result in poor alignment due to sparseness. Because MassCascade and Obiwarp store MS data differently, binned grids are stored temporarily in Obiwarp-compatible ASCII ‘lmat’ format for every sample and reference. Lmat files can subsequently be used as input for Obiwarp, spawned in parallel in separate threads. Obiwarp returns a vector or adjusted time values of equal length to the number of bins in the time domain of the input grid. Nearest-point linear interpolation is used to translate these adjusted time values into corrections for the unbinned features in the original feature collection (Figure 2.9b).

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

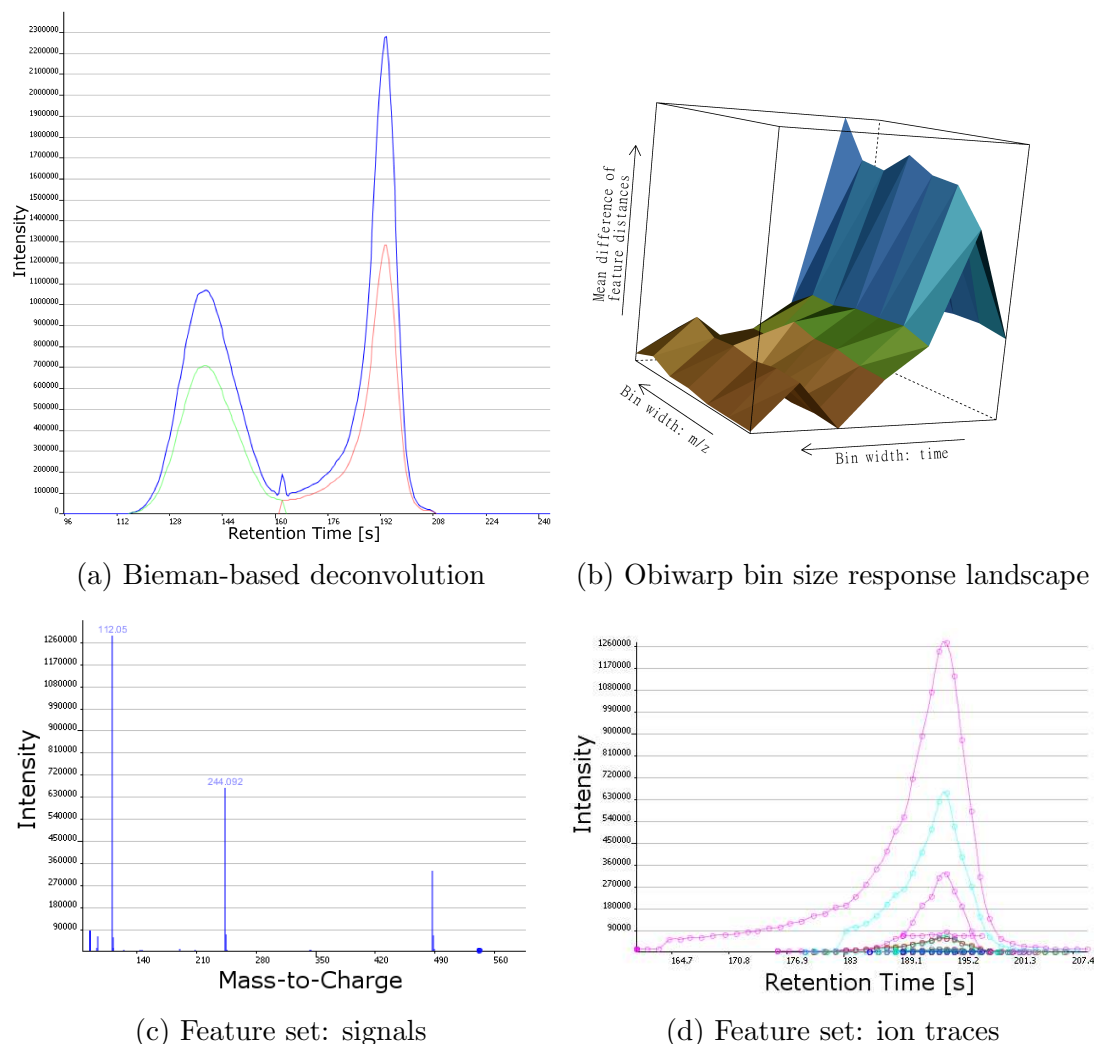


Figure 2.9: Illustration of deconvolution, alignment, and feature set methods. (a) Example of Bieman-based deconvolution. The outer ion trace (blue, scaled up) represents the chromatogram of a single ion. The green and red traces result from the deconvolution process. The original ion trace has been truncated at its local minima. (b) Response landscape of the mean difference of target-to-reference distances before ( $\Delta t$ ) and after ( $\Delta t'$ ) alignment with Obiwrap:  $\overline{\Delta T} = \frac{1}{n} \sum_{i=0}^n (|\Delta t'_i| - |\Delta t_i|)$ , where  $n$  equals the number of aligned features; the times  $\Delta t$  and  $\Delta t'$  correspond to the differences in retention time  $r_t$ :  $\Delta t^{(l)} = t_r^S - t_r^R$  with  $S$  and  $R$  being the target and reference. Bin widths in the time and  $m/z$  domain have been increased in steps of 0.5 from 0.1 to 3.0. Lower  $\overline{\Delta T}$  values indicate better overall alignment. The graph shows that best alignment results are achieved with larger bin sizes for the time domain, which should approximate the time interval between scans. Obiwrap was run with default settings. (c) Spectrum and (d) ion traces of a feature set representing Cytidine ( $[M+H]^+ = 244.092$ ) grouped *via* cosine similarity.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

Atypical for LC-MS software, MassCascade does not compile fragmentation mass spectra by default. One or more fragmentation spectra, i.e. MS<sup>2+</sup>, are typically generated for selected precursor ions. For these ions, every MS<sup>1</sup> scan results in one or more MS<sup>2+</sup> scans that are recorded. These can be collapsed into a single representative fragmentation spectrum similar to ion extraction in MS<sup>1</sup> (see paragraph 2.3.1). Representative fragmentation spectra of MS depth  $n$  are subsequently assigned to the extracted feature (ion trace) of the precursor ion in recursive fashion, e.g. for structure fragment assignment. Note that only selected features with experimentally acquired MS<sup>2+</sup> scans can have associated consensus spectra. It should be noted that collapsed fragmentation spectra are consensus spectra that only contain signals present in all fragmentation scans of a parent ion. This process has been kept separate in line with the principle of modularity.

A collection of features is highly redundant by itself. Multiple signals can result from a single molecular compound. Cosine similarity and a modified Bieman approach have been implemented to group related features into feature sets (see chapter 2.2).

The cosine similarity function creates a pairwise similarity matrix from all features of a sample. To reduce memory load, only features within a pre-defined time interval are considered simultaneously. Non-overlapping ion traces are excluded by default. The spectral vectors, i.e. the features' ion trace intensity vectors  $(\vec{I}_a, \vec{I}_b)$ , are used to calculate their similarity (angle  $\theta$ ) using the dot product and magnitude:

$$\text{sim} = \cos(\theta) = \frac{\vec{I}_a \cdot \vec{I}_b}{\|\vec{I}_a\| \|\vec{I}_b\|} \quad \text{sim} \in \{0 \leq \text{sim} \leq 1\} \quad (2.17a)$$

Correlated features show similarities greater than 0.95. Any features below the defined similarity threshold are not clustered together in a single feature set (Figure 2.9c and 2.9d). If multiple features exist that have overlapping but gradually shifting ion traces, the algorithm will follow these shifts in the time domain to group the features together. Higher similarity thresholds restrict that behaviour automatically.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

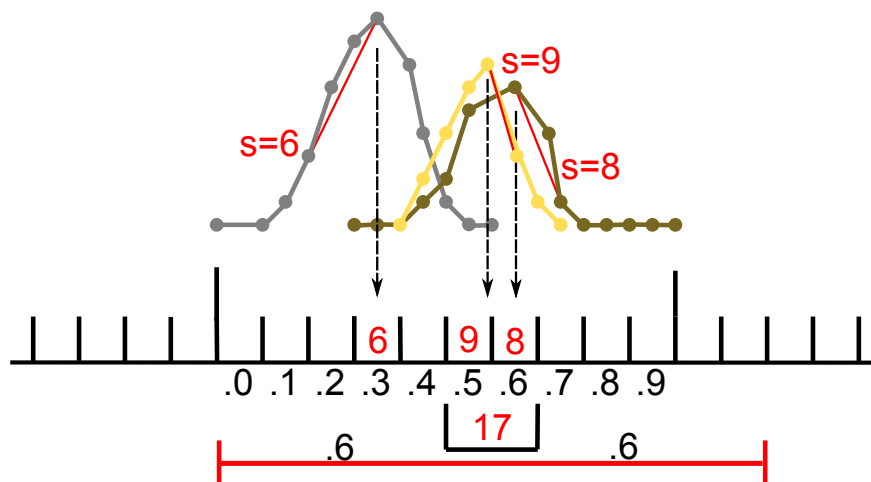


Figure 2.10: Illustration of the modified Bieman algorithm to group related features. Overlapping features are binned in the time domain using a bin size of 10. Three features are shown in gray, yellow, and brown. Individual features are binned by their sharpness value  $s$ , i.e. a value describing the steepest slope from their apex to a neighbouring data point (indicated by the red line). To estimate the uncertainty range, the number of bins is divided by the sum of values of populated neighbouring bins:  $\frac{10}{9+8} \approx 0.6$ . If no bin within the uncertainty range contains a higher sharpness value than the maximum value in the bins that gave rise to the uncertainty range, features captured in those bins are believed to be correlated (the yellow and brown peaks). The figure is adapted from Stein *et al.*<sup>[113]</sup>.

The modified Bieman approach uses a shape-based approach described by Stein *et al.*<sup>[113]</sup>, Figure 2.10. The granularity of the clustering can be adjusted by the number of bins. A single bin is defined by a given length (parameter A) divided by the given number of bins (parameter B). For this method, the length should approximate the distance of two adjacent scans for fine resolution. Higher length values will yield larger feature sets with less similar features grouped together.

### 2.3.3 Data Post-Processing

The annotation and identification of features or feature sets forms the last part of an LC-MS<sup>n</sup> analysis. The identification of metabolites is discussed in detail in the next chapter. This section deals with the annotation and low-confidence identification only. MassCascade stores meta data extracted from raw files, such as a

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

sample’s ion mode, internally to simplify usage. Thus, the number of parameters for annotation methods has been reduced to the greatest degree possible because information about MS<sup>n</sup> levels or ion modes does not need explicit input.

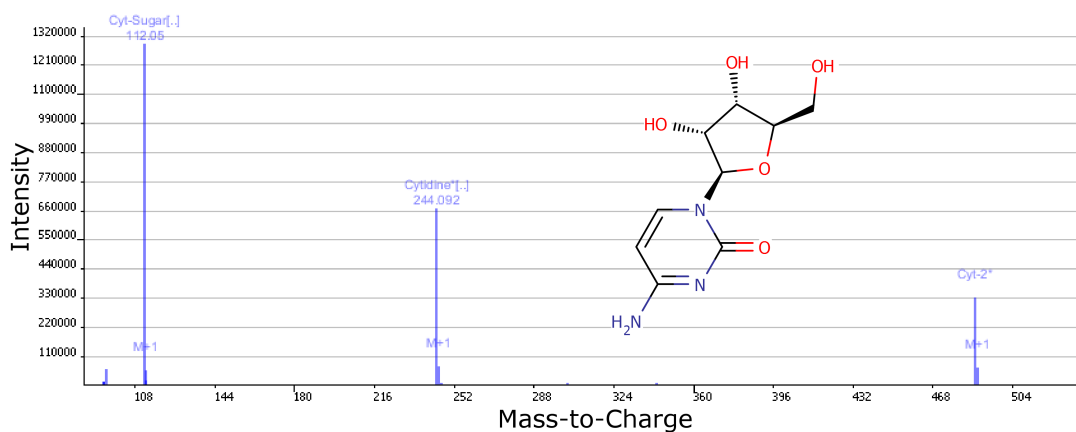
Adducts, neutral losses, and any other relations that are based on differences of  $m/z$  signals within a feature set can be annotated: pairwise signals for which a relationship evaluates to ‘true’ are annotated with the corresponding feature identifiers and type of relationship (Figure 2.11a). For example, a  $m/z$  difference of 21.9819 represents a replacement of a hydrogen atom with a sodium atom in positive ion mode:  $[M+H]^+ \rightarrow [M+Na]^+$ . Feature sets can be queried for any two-column list of annotations and  $m/z$  differences. The  $m/z$  differences are matched for every feature pair within a given ppm tolerance. Comprehensive adduct and neutral loss lists have been published that can be used as template<sup>[110]</sup>.

Isotopes are detected using a combinatorial approach with a quality check for final assignment based on a heuristic model. The average isotope difference is set to  $\Delta mz = 1.0033$ <sup>[234]</sup>. Isotopic signals can be detected up to a distance of three, i.e.  $[M+3]$ . Abundances of single-step isotopic signals with distances greater than three are considered to be of too low intensity to be relevant: with an isotopic abundance of  $^{13}\text{C} = 1.10$ ,  $\text{C}_{66}$  would result in a signal at  $[M+4]$  that equals 1% of the intensity of the main signal. Elements that have ‘irregular’ isotopic patterns, e.g.  $^{32}\text{S} = 95.02\%$ ,  $^{33}\text{S} = 0.75\%$ ,  $^{34}\text{S} = 4.21\%$ , are not fully detected. A  $m/z$  tolerance can be defined in ppm to account for mass defects. The algorithm sorts all features of a feature set in ascending  $m/z$  order before traversing the list in recursive fashion, pairwise matching all features lower in the list to the current index. If a  $m/z$  difference equals the theoretical isotope difference, the relative intensity of the isotope signal, i.e. its intensity divided by the main signal intensity, is matched to the theoretical relative intensity of an isotopic signal within the corresponding  $m/z$  range in which the signal falls. A positive match results in isotope annotation.

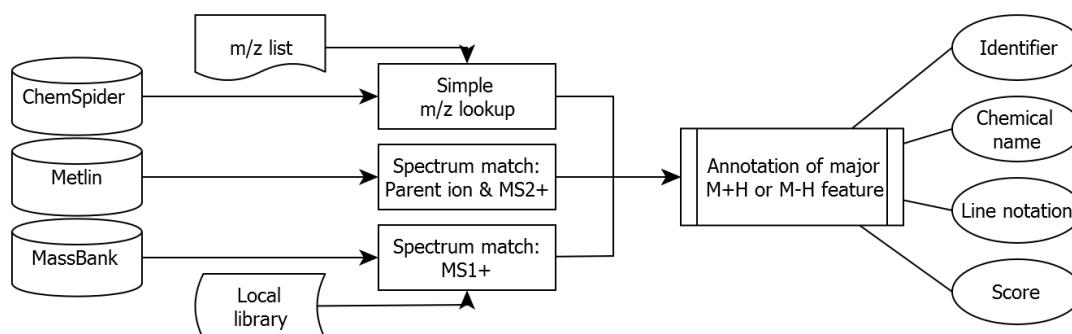
The heuristic model for the theoretical relative intensity of  $m/z$  ranges was derived in the following way: small molecules were downloaded in bulk in SD-File format from the metabolomics databases HMDB<sup>[235]</sup>, Golm<sup>[236]</sup>, and MassBank<sup>[237]</sup>. Molecular formulas were extracted and filtered for subsets contain-



## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS



(a) Isotope and identity annotations for Cytidine



(b) Diagram of feature identification methods and their mode of action

Figure 2.11: Illustration of annotation and identification methods. (a) Feature set (compound spectrum) of Cytidine (structure shown) with annotations for isotopes as perceived by MassCascade. Relations such as the Cytidine cluster (Cyt-2) or its sugar-removed fragment (Cyt-Sugar) were annotated through a list with  $m/z$  differences. (b) Diagram of feature or feature set identification methods including their mode of action. ChemSpider, Metlin, and MassBank are online databases. Reference library and  $m/z$  list queries are local operations. Whereas a ‘Spectrum match: MS<sup>1+</sup>’ matches the feature set at MS<sup>1</sup> as well as relevant fragmentation spectra (MS<sup>2+</sup>), the ‘Parent ion & MS<sup>2+</sup>’ match by Metlin only matches a single feature in MS<sup>1</sup> and subsequently proceeds to its fragmentation spectra. Matching features are annotated with a small molecule’s database identifier, chemical name, line notation (SMILES or InChI if provided), and query score.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

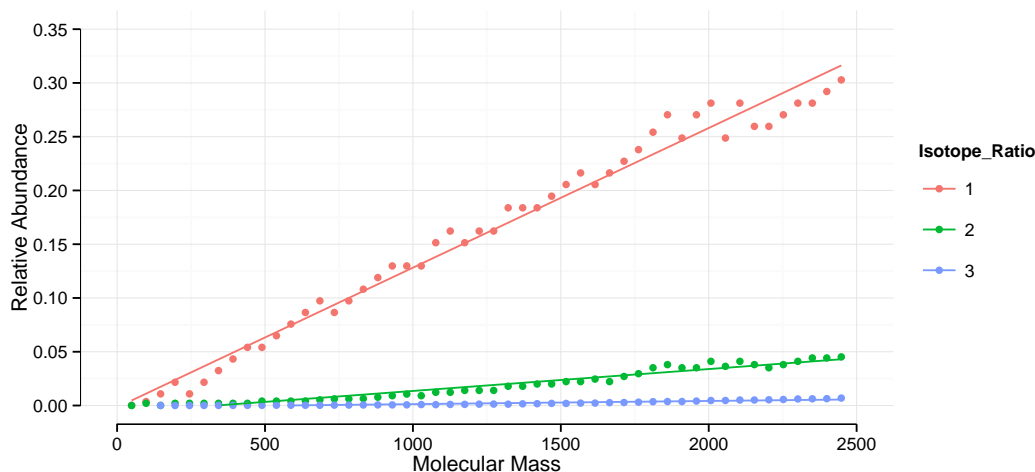


Figure 2.12: Linear regression for molecular masses vs. isotope abundances. Data series for isotope ratios 1 (red), 2 (green), and 3 (blue) are shown. The molecular masses and theoretical isotope patterns were calculated from the molecule collections of HMDB, Golm, and MassBank.

ing only the elements CHNOPS + F + Cl. Bromide was excluded because of its atypical isotopic abundance pattern,  $^{79}\text{Br} = 50.69\%$ ,  $^{81}\text{Br} = 49.31\%$ , which would significantly skew the relative intensity distribution for isotopic distances. Sulfur,  $^{32}\text{S} = 94.93\%$ ,  $^{33}\text{S} = 0.76\%$ ,  $^{34}\text{S} = 4.29\%$ , and Chlorine,  $^{35}\text{Cl} = 75.78\%$ ,  $^{37}\text{Cl} = 24.22\%$ , did not impact significantly on the intensity distribution. Their isotopic abundance patterns' 'irregularities' were absorbed in the overall measurement error. The representative, non-unique set contained 47,849 molecular formulas. The theoretical isotopic envelopes were calculated for the whole set and a linear regression applied for the relative intensities of every isotopic ratio (Figure 2.12). The resulting linear formulas of form  $I = a * m/z + b$  describe the deduced intensity  $I$  for any  $m/z$  with the coefficients  $a$  and  $b$ .

Putative feature or feature set identifications, i.e. metabolite annotations, are facilitated by methods for direct ion identification, local library queries, and web-based searches (Figure 2.11b). Direct ion identification refers to simple  $m/z$  lookup in a reference list that can directly be fed into MassCascade. The reference list has to contain exact molecular masses and molecule labels. The molecular masses are automatically (de-)protonated based on the ion mode.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

Local library queries depend on the creation of a `ReferenceContainer` instance, containing `ReferenceSpectrum` instances. The reference spectra have a  $m/z$ -intensity list for each set MS level and compound information such as the name, line notation, molecular mass, and formula. Fragmentation spectra are traversed if defined. Local queries are dramatically faster than web-based searches.

Web-based searches include the databases ChemSpider<sup>[186]</sup>, MassBank<sup>[237]</sup>, and Metlin<sup>[238]</sup>. ChemSpider and Metlin queries require a security token issued by the individual websites. For the databank ChemSpider a subset of databases can be selected for simple  $m/z$  lookups from the total list of databases available. The MassBank and Metlin integration offers feature set (MS<sup>1</sup>) and fragmentation spectra matching (MS<sup>2+</sup>) in addition to simple  $m/z$  lookups, resulting in more reliable feature annotation. Input parameters correspond to the web interfaces. Results from local or web-based queries are assigned to the respective features or features sets including the chemical name, line notation (SMILES or InChI), query score, and database identifier.

### 2.4 MassCascade for KNIME

A plug-in for the workflow environment KNIME has been developed. The plug-in MassCascade-KNIME wraps the functionality of the MassCascade core library and exposes the functions as graphical *nodes*. A *node* can be considered an atomic unit of interaction. *Manipulator nodes* take one to many inputs and produce one to many outputs. *Source nodes* take no input from other nodes whereas *sink nodes* do not pass any output further. Similar to a scripted pipeline, individual functions, i.e. nodes, can be inserted or deleted at any point in the workflow.

#### 2.4.1 Structure

The plug-in serves as interface between the open-source platform KNIME and the MassCascade library. It manages node-user and node-node interactions and delegates jobs to the task classes of its back-end. These jobs are grouped into three

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

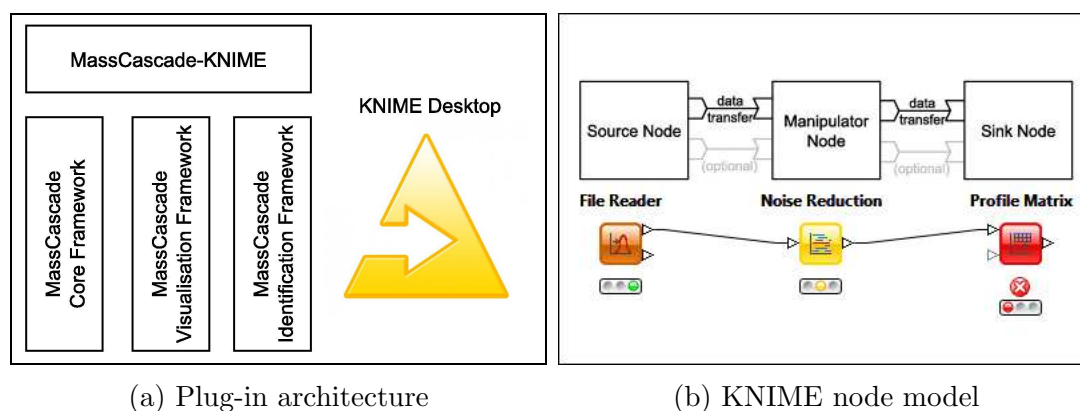


Figure 2.13: Overview of the MassCascade-KNIME architecture and KNIME’s node model. (a) The plug-in forms part of the KNIME Desktop and serves as interface between the three frameworks of the MassCascade library and the KNIME workbench. (b) The adapted KNIME node model distinguishes between *source*, *sink*, and *manipulator nodes*. These node types have distinct input and output ports and facilitate intuitive usage of the plug-in. The ‘traffic light’ color scheme indicates the state of a node: Executed nodes are *green*, configured nodes are *yellow*, and unconfigured nodes are *red*.

categories based on their purpose: jobs that address the core framework (data processing), the visualisation framework (plots and charts), or the identification framework (signal or spectrum identification, rationalisation of results) (Figure 2.13a). Each category is handled differently in the plug-in (section 2.4.3).

In contrast to the internal design, nodes are grouped by the kind of interaction they provide: Source nodes read data from files or convert external tabular data to a format that can be processed by the plug-in, manipulator nodes change the data, sink nodes write data to files or convert internal tabular data to generic data types, and visualisation nodes provide views of data in tabular or graphical form (Figure 2.13b).

Underneath the nodes, data processing jobs are executed concurrently using Java’s<sup>®</sup> concurrency framework<sup>[229]</sup>. Each thread deals with a single sample, i.e. reads the data container, applies a function, and writes the data out into a new container. The number of threads available depends on the global configuration of KNIME. Raw data is serialized to disk by the MassCascade library in a temporary directory specified by the user whereas access pointers to the

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

raw data are stored through KNIME's persistence framework. Employing this split serialisation scheme guarantees efficient data handling: The KNIME plug-in only deals with small XML-based files that contain pointers to large data chunks. These small files can be modified and swapped back and forth between nodes at low cost, e.g. memory. For operations on the LC-MS<sup>n</sup> data, the files are converted to their respective container types and passed on to the main task class. The main task class reads the data on demand and returns a new container, as described in section 2.2. The returned container is again serialised to XML for persistence.

### 2.4.2 Node Types

Nodes of different types can be concatenated to build a workflow as depicted in Figure 2.14. Every MassCascade-KNIME workflow must start with a source node and typically ends with a sink node to output results. Manipulator nodes in between a source and sink node are for data processing. Visualisation nodes can be employed at any step in the workflow to inspect the data at that node. In contrast to all other node types, visualisation node types do not require any input parameters.

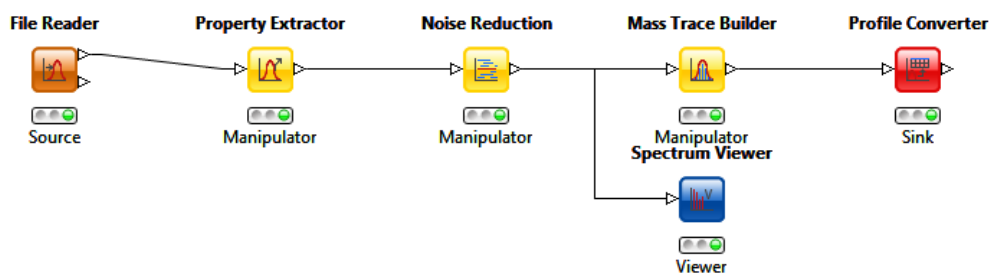


Figure 2.14: Typical MassCascade-KNIME workflow demonstrating how different node types can be used in combination. LC-MS<sup>n</sup> files are read in *via* the *File Reader* node before the extraction of meta information. Following, random noise is eliminated and features are extracted. The features are converted to a generic tabular KNIME format that can be used with generic KNIME nodes. The *Spectrum Viewer* allows interactive inspection of the noise-reduced samples.

### 2.4.3 Node Interactions

Two types of node interactions are distinguished: Node-node and node-user interactions. Both types of interactions are defined by the plug-in. To render the node coding process more efficient, abstract custom models have been built on top of the four key parts of a default KNIME node model (see section 1.2.4). These abstract classes are extended by all MassCascade-KNIME nodes, enforcing standardised behaviour across all interactions levels and reducing time spend developing nodes (Figure 2.15). Because of the common base classes, any new node implementation requires only the implementation of a model and – if required – view, in addition to the definition of node parameters.

#### Node-Node Interactions

Node-node interactions are based on KNIME's tabular data model (see section 1.2.4). Three distinct data cell types have been developed to represent MS, profile, and spectrum data containers of the core library (see section 2.2). Instances of these cell types are serialized to disk for persistence by the KNIME environment. Node-node interactions are only possible when a required input cell type of one node matches an output cell type of a preceding node. This compatibility of cell types is enforced by an automated mechanism in the *DefaultModel* class. Input or output of multiple data tables must be specifically provided in the model implementation but follows the same principle. MassCascade cells types depict their content in KNIME tables: MS data cells sketch a total ion chromatogram, profile data cells depict a histogram of time bins (ascending) to number of binned profiles, and spectrum data cells show a plot of time versus  $m/z$  values, where every point represents a found profile.

#### Node-User Interactions

Node-user interactions are provided through node dialogues, depictions, input-output tables, and data views. MassCascade-KNIME provides interactive parameter dialogues with parameter validation, default values, and auto-configuration

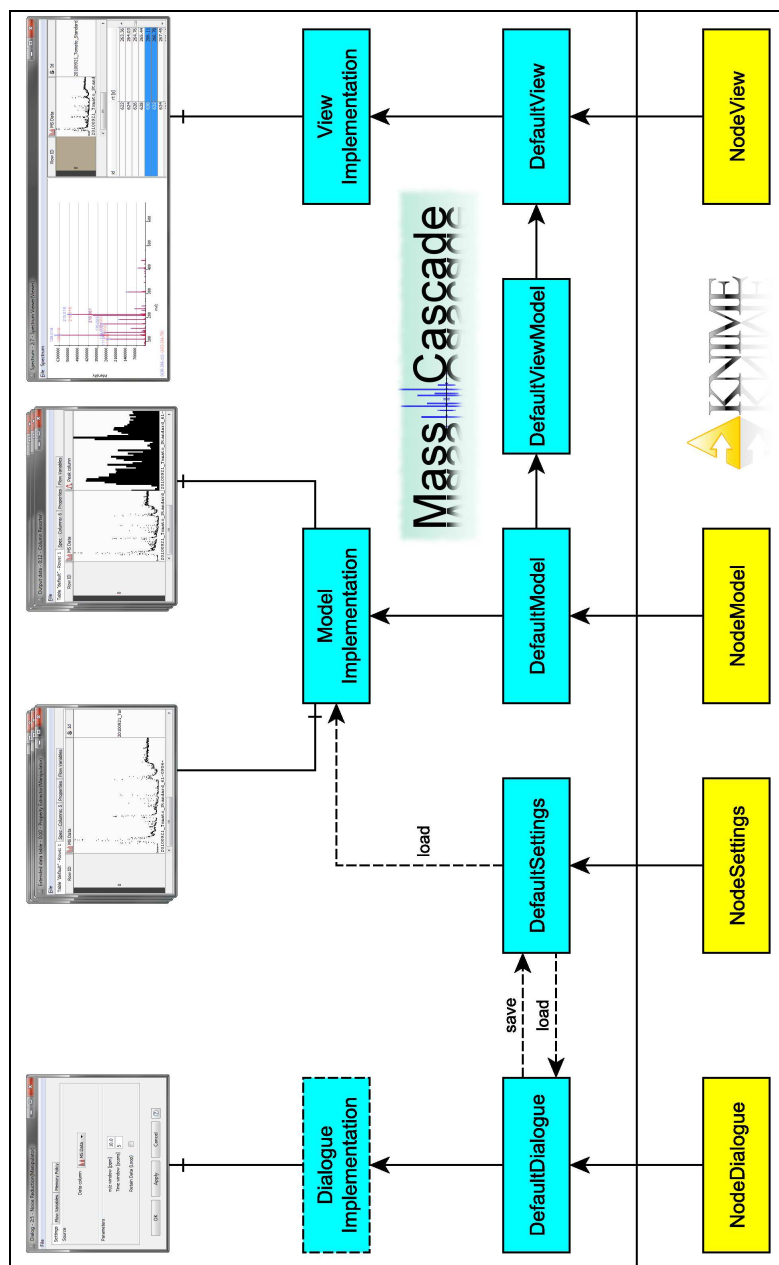
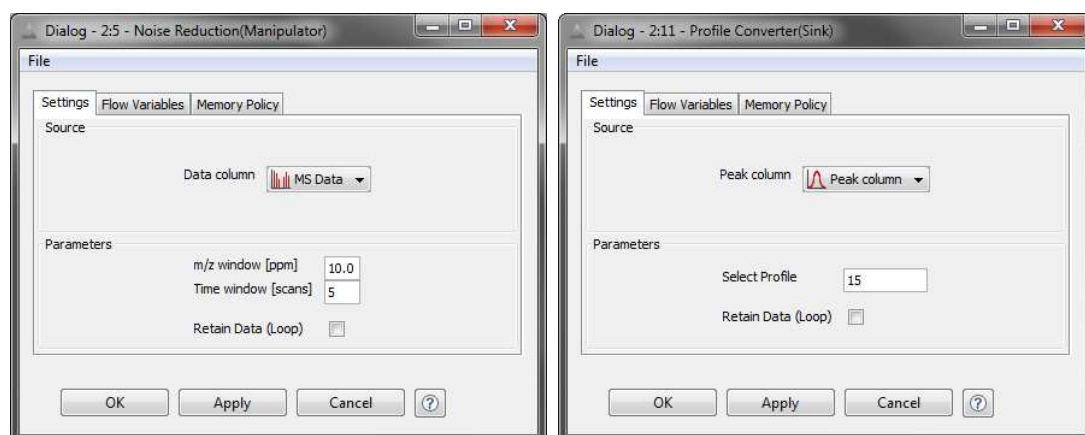


Figure 2.15: Schematic of the node architecture and interactions. The generic KNIME classes have been extended to create abstract default classes to simplify the development process. The *DefaultSettings* class links the configuration dialogue and node model, saving and loading all node parameters. The *DefaultDialog* class is a template that builds configuration dialogues on-the-fly based on pre-defined building blocks. It loads/saves node parameters to/from the default settings class. The *DefaultModel* class provides parameter validation and handles node execution. It takes a set of input tables and produces a set of output tables. The *DefaultView* class visualises data exposed through the *DefaultViewModel* class in the form of graphical charts.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---



(a) *Noise Reduction* node dialogue

(b) *Profile Converter* node dialogue

Figure 2.16: Configuration dialogues of the (a) *Noise Reduction* and (b) *Profile Converter* nodes. The dialogues take a set of input parameters in the ‘Parameters’ section and define the required input cell type in the ‘Source’ section of the tab.

(Figure 2.16). Erroneous parameters are not accepted and render the node disabled. Node dialogues are loaded with default parameters. Auto-configuration checks if the input table(s) contain the data types required. The outcome of the node configuration is reflected in the traffic light representation of the nodes’ execution status.

Data views visualise data cells of a particular type. Execution of the visualisation nodes does not change data in the respective data cells. All visualisation nodes follow the same layout: input data cells are listed on the right of the visualisation window, optionally with an added section for further listings for loading data. The main part of the window is taken up by an interactive two dimensional plot. Depending on the individual node, plots offer different data representations, accessible *via* the file menu (Figure 2.17).



## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

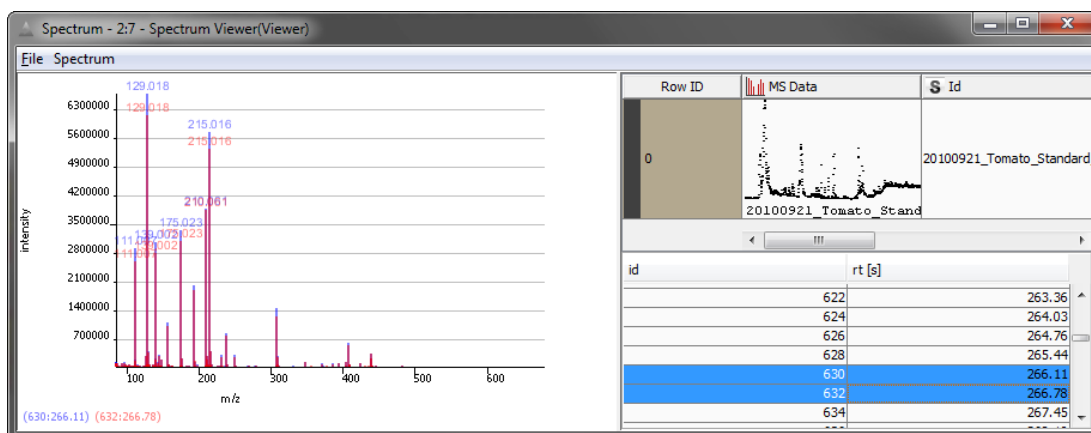


Figure 2.17: Demonstration of the *Spectrum Viewer* node. The central area of the window shows two spectra selected *via* the spectrum list in the bottom right. The data cells for the spectrum list are shown in the top right.

## 2.5 Evaluation

### 2.5.1 Spectral Fingerprinting of Tomato Samples

Metabolomic fingerprinting is one of the key applications of LC-MS<sup>n</sup>. Metabolomic fingerprinting is based on a sample by feature matrix, where features have reliably been found across samples. These features are believed to be characteristic for the samples and can thus be used for multivariate clustering and subsequent generation of predictive models. Fingerprinting relies on the extraction of real features, i.e.  $m/z$ -intensity value pairs that represent ions stemming from metabolites, and on correct alignment of features across samples to generate the feature matrix. The quality of the resulting matrix depends heavily on the feature extraction and alignment process, as well as on the overall data pre-processing. Insufficient or incorrect data pre-processing could introduce noise to the matrix in the form of irrelevant features, e.g.  $m/z$ -intensity value pairs that stem from background ions or impurities, or incorrect features, e.g. value pairs that do not exist or are not present in all samples indicated. This matrix would show a higher level of unexplainable variance and could lead to wrong clustering.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

MassCascade was tested for correct fingerprinting using a data set of aliquots of a pooled standard tomato control with known clustering. The test allows for exploration of various aspects of the tool, including feature extraction, deconvolution, and alignment, and provides a statistical measures for the quality of the generated feature matrix.

### 2.5.2 Materials and Methods

The data set was provided by the Syngenta AG. The data was acquired using aliquots of a polar extract from wild-type tomato (*Solanum lycopersicum*), which was used as pooled in-house control. The 113 samples were used as controls in a long-term study at the Syngenta AG (section 2.1). Approximately 30 mg of dried tomato tissue was extracted with ethanol:water 20:80 (v/v) and diluted 10:1 (v/v) with water for injection. A mixture of citric acid-d4 [CID 16213286], L-alanine-d4 [CID 12205373], glutamic acid-d5 [CID 56845948], and L-phenyl-alanine-d5 [CID 13000995] was added for internal calibration. The samples were measured on a Thermo Velos Orbitrap coupled to a Waters Acquity UPLC with a HSS T3 150 x 2 mm, 1.7  $\mu$ m, Acquity UPLC column. The solvents used for the assay consisted of 0.2% formic acid (solvent A) and 98/2/0.2 acetonitrile/water/formic acid (solvent B). The gradient started at 100% A (hold 2.5 minutes) at 0.25 mL/min followed by a ramp to 10% B after 7.5 minutes increasing the flow rate of 0.4 mL/min; then to 100% B after 10 minutes, hold 2 minutes, before equilibrating back to starting conditions after 18 minutes. The samples were detected in positive ESI mode at a resolving power of 30,000 FWHM with a scan range from 85-900 Da. MS<sup>2</sup> spectra were obtained in a data dependent manner: The two most intense mass spectral peaks detected in each scan were fragmented to give MS<sup>2</sup> spectra. Full scan data was acquired in profile mode, MS<sup>2</sup> spectra were acquired in centroid mode.

The data files were converted from Thermo Scientific's *RAW* format to *mzML* using ProteoWizard v2.2<sup>[230]</sup> with vendor-based peak picking enabled for MS<sup>1</sup>. The data were processed using the MassCascade plug-in for KNIME, following the steps outlined in Figure 2.18. Feature extraction was carried out with 10

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

September		October		January		March	
Date	Count	Date	Count	Date	Count	Date	Count
2010-09-17	4	2010-10-04	1	2011-01-18	6	2011-03-22	4
2010-09-18	9	2010-10-05	8	2011-01-19	7		
2010-09-20	7	2010-10-06	15				
2010-09-21	14	2010-10-07	10				
2010-09-22	9	2010-10-08	15				
		2010-10-09	4				

Table 2.1: Overview of the number of wild-type tomato samples measured by date. The samples were measured over a total of four months.

ppm mass accuracy. A minimum width of six scans and a minimum intensity of 100,000 units was used for thresholding of features. In addition, a Durbin-Watson statistic was used to filter out features that were assigned a score above the third quartile of the normal score distribution taken from all features, here  $Q_3^{dw} = 2.38$ . Deconvolution was applied with a signal to noise ratio of two using a modified Bieman algorithm. Obiwrap was used for cross-sample alignment with default parameters and aligned features subsequently filtered and selected for presence of isotopic peaks. Features associated with common interferents were removed.

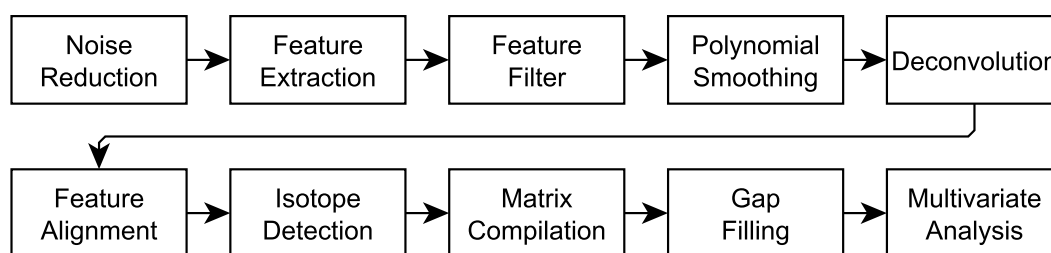


Figure 2.18: Flow diagram depicting LC-MS data processing steps for metabolomic fingerprinting. After random noise reduction, features are extracted and filtered using a statistic describing expected chromatographic behaviour. The filtered features are smoothed in the time domain and deconvoluted to resolve overlapping isobaric features. Found features are matched across all samples of a group, filtered for presence of isotopic peaks for selectivity, and written out in feature matrix. Subsequently, intensity gaps in the matrix are filled before multivariate analysis.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

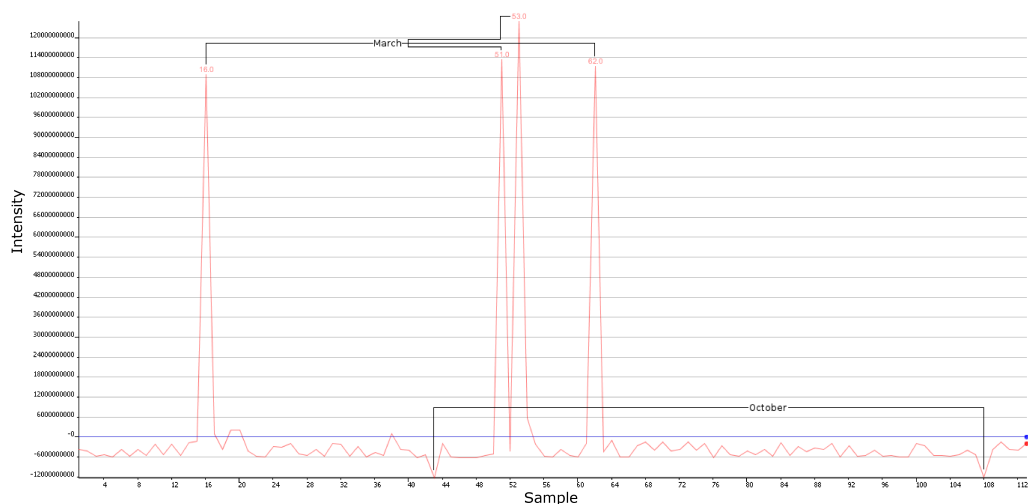
The aligned features were written out in a feature matrix and gap filling was applied: gaps were filled with intensity values from the baseline of the individual samples at the features' retention times and  $m/z$  values. If no baseline value could be found, gaps were assigned a default intensity of 10,000 units. A total of 11 result matrices were compiled with allowed missingness ranging from 0-100% in 10% increments. The matrices were subjected to multivariate analysis in the statistical computing environment R.

### 2.5.3 Results

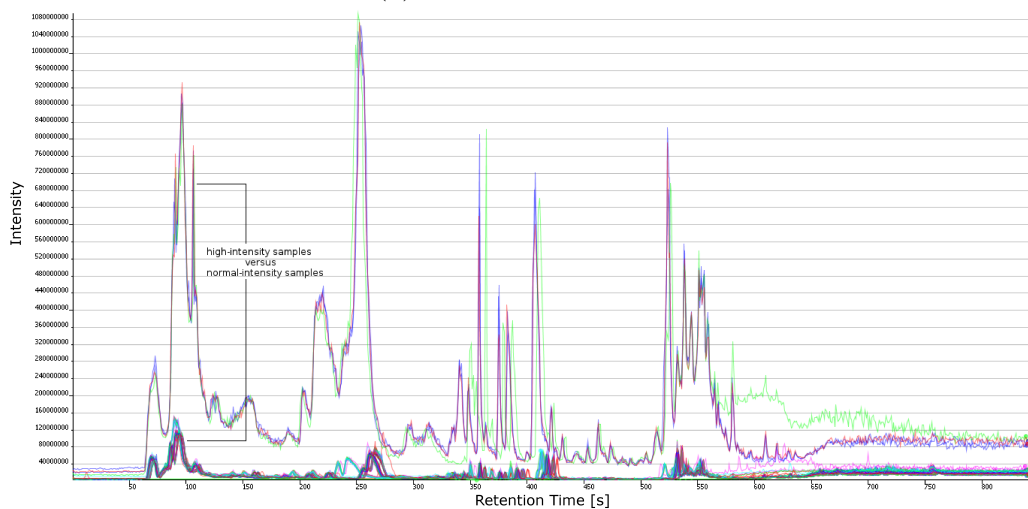
Initial inspection of the total ion chromatograms of all samples revealed two failed injections for samples 2010-10-09\_12 and 2010-10-09\_13. Both samples contain only baseline noise with no apparent features and were removed from analysis. Cross-sample comparison of mean-centered total ion currents showed significantly higher intensities for samples 2011-0322\_01, 2011-03-22\_02, 2011-03-22\_03, and 2011-03-22\_04, likely due to wrong sample preparation or over-injection (Figure 2.19). All four March samples were flagged as outliers and kept.

Sample 2010-09-21\_61 was chosen as reference for the feature-based alignment with Obiwarped based on its representative total ion current and shape of its total ion chromatogram, minimising overall time shift. The result of the non-linear alignment is depicted in Figure 2.20a. The stability of the column measured by sample to reference time deviation appears to be dependent on the acquisition date. Time shifts of features grow bigger with an increase in the time gap between sample and reference. Samples acquired on the same date show the least variation in time shifts, followed by variation in between groups acquired in the same month. This was to be expected. Even though the same instrument was used for all measurements, it was not used exclusively for the study. Other experiments were run on the Orbitrap in between the months. Thus, column modifications such as degradation or insufficient equilibration become more apparent over time. All samples show uncorrelated shifts before about 80 s and after 750 s, coinciding with the dead time of the system and end point of the solvent gradient respectively.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS



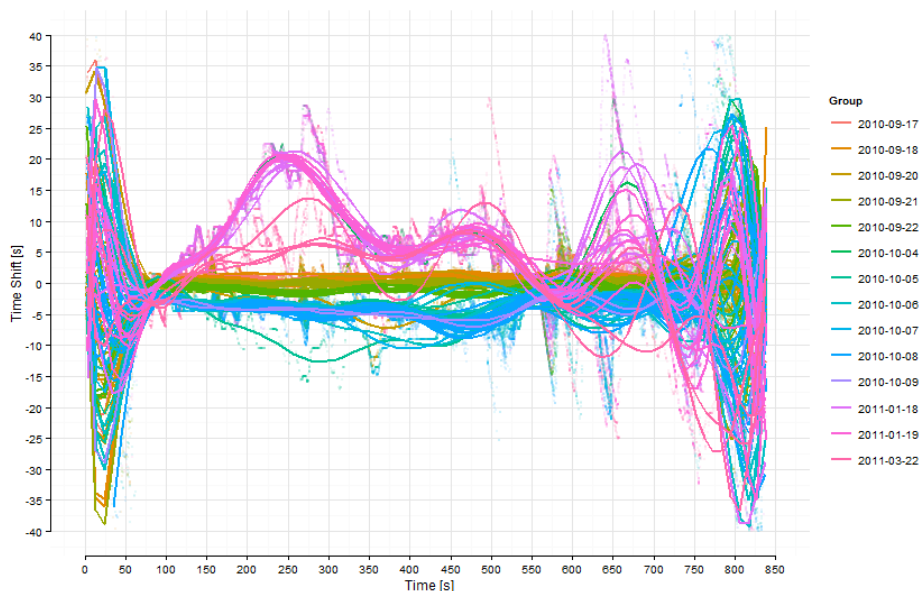
(a) Total ion currents



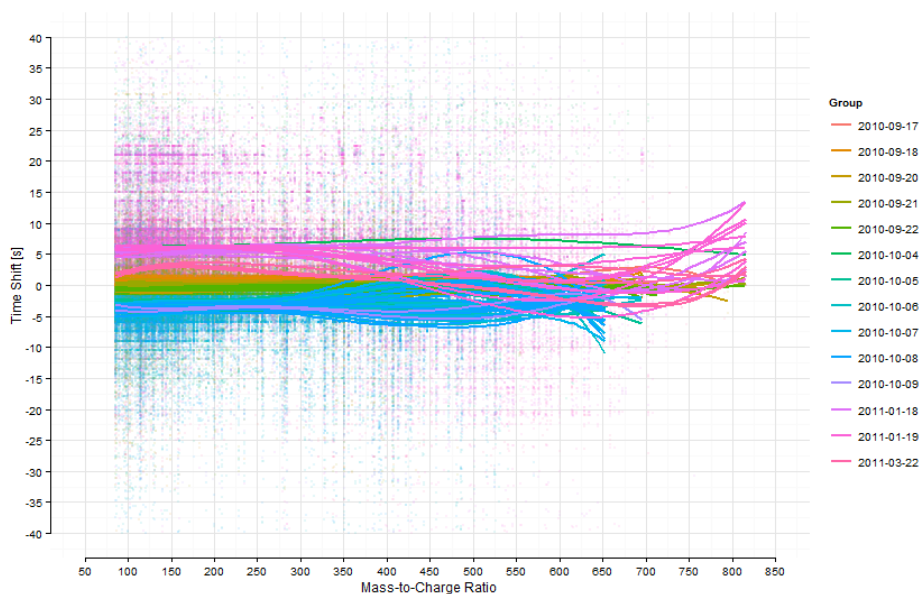
(b) Total ion chromatograms

Figure 2.19: Summary of the sample intensity domain. (a) The mean-centered total ion current for each sample is shown in red. The blue line indicates zero intensity. Samples 16, 51, 53, and 62 represent the four March samples and show significant deviation from the trend. Samples 43 and 108 refer to 2010-10-09\_12 and 2010-10-09\_13 and contain baseline noise only, explaining the strong negative deviation. (b) Overlay of all total ion chromatograms coloured by date. All chromatograms have similar shape. The four March samples are clearly distinct with dramatically higher intensities. The samples show slight time deviation and increasing baseline drift from 570 s.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS



(a) Time - shift regression



(b)  $m/z$  - shift regression

Figure 2.20: Summary of the feature alignment. The data points and regression curves are coloured by acquisition date. (a) Plot of the elution time versus time shift, i.e. the time difference before and after alignment. A polynomial regression fit was used. The regression curves are coloured by acquisition date. Samples measured on the same date cluster together. Shifts before about 80 s and after 750 s appear to be uncorrelated. (b) Plot of the  $m/z$  values versus time shift with a polynomial regression fit. The linear trends show that the time shift is almost constant for all  $m/z$  values, indicating no correlation between mass-to-charge ratio and time shift.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

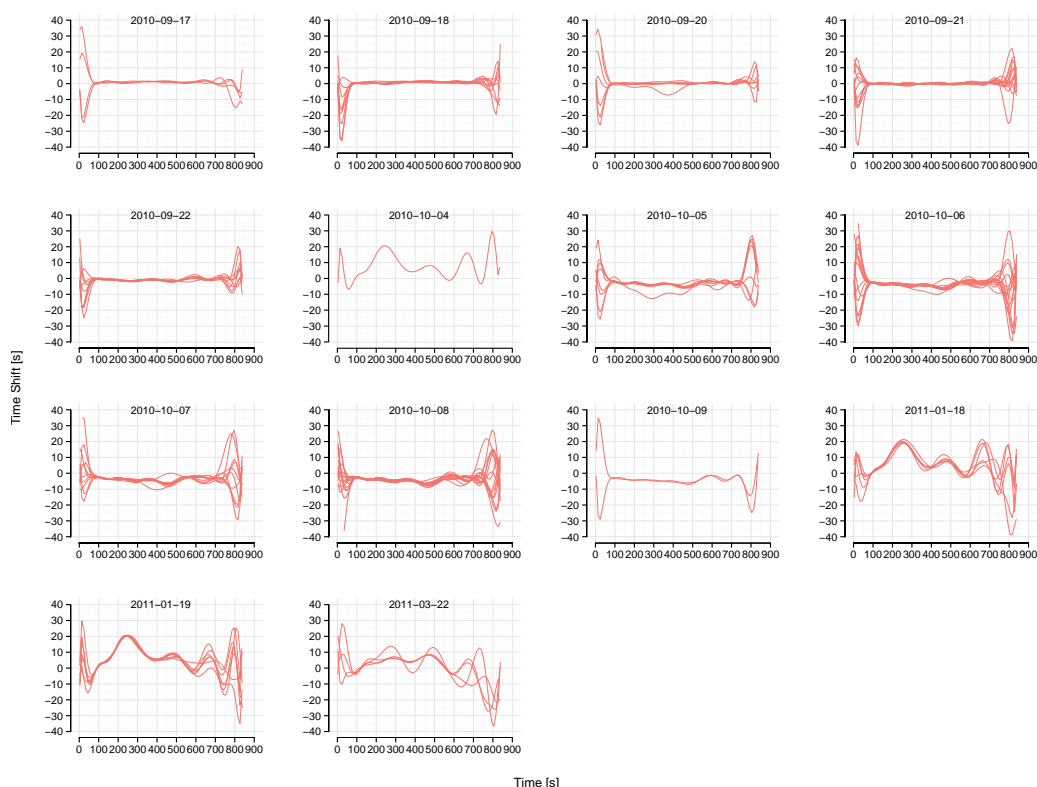


Figure 2.21: Overview of the feature alignment groups. The matrix shows the individual regression plots for all sample groups. The elution time versus time shift is plotted with a polynomial regression fit.

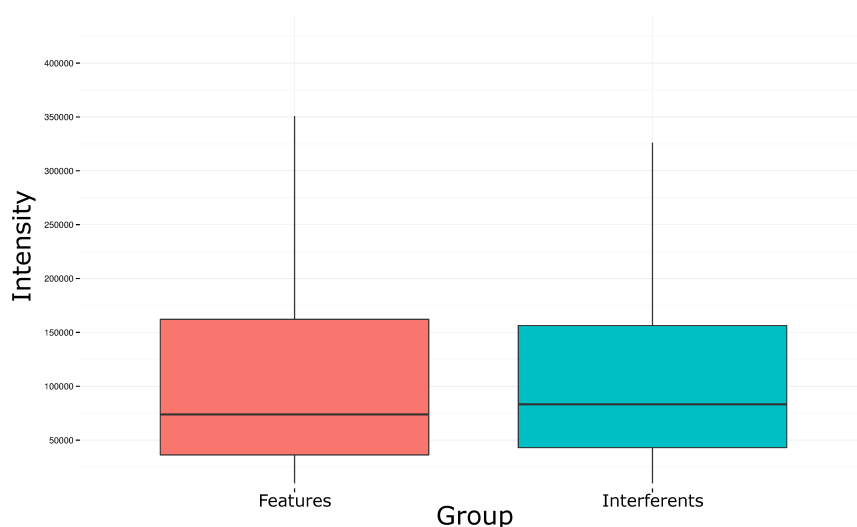
The individual groups are shown in Figure 2.21. The mass-to-charge values seem to be uncorrelated to time shifts and have no impact on the alignment (Figure 2.20b).

Removal of common interferences (see appendix section 5) in LC-ESI-MS reduced the number of features by an average of 212 per sample. The matched features show no significant differences in their distributions compared to the remaining features (Figure 2.22a). Table 2.2 lists the top five most frequently found contaminants. N-methyl-2-pyrrolidone, oleamide, and polyethylene glycol are believed to come from laboratory equipment used in sample preparation. The acetonitrile/acetic acid species stems from the mobile phase. Its putative sodium adduct shows a higher mass deviation of 19.0 ppm. The mass chromatograms of both species were compared to verify that they are related. They co-elute at various

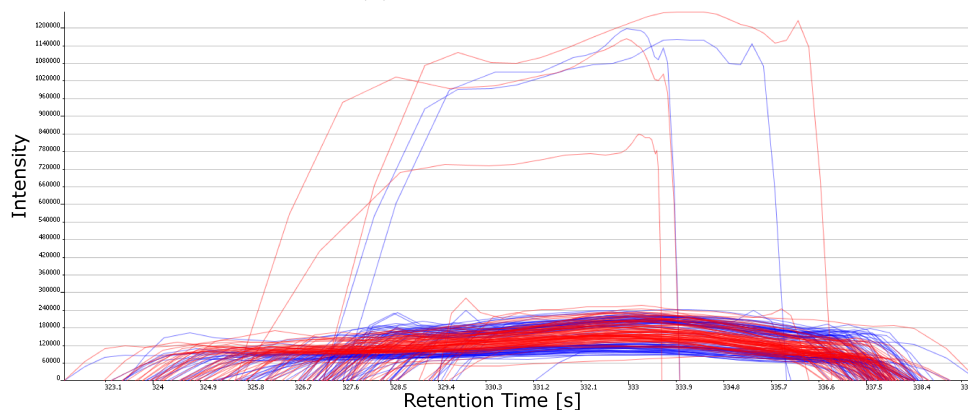
## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

time points during the runs with the acetonitrile/acetic acid species present in all samples and the sodium adduct absent in up to 11 samples, in consistence with the assumption that the adduct is detected less often than the main species. Comparison of the shapes of the feature mass chromatograms at 333 s support this finding of a main species to sodium adduct relationship (Figure 2.22b).



(a) Intensity boxplot



(b) Mass chromatograms of two interferents

Figure 2.22: Summary of the interferents analysis. (a) Boxplot of the intensity distribution of features versus filtered out interferents. No outliers are shown. The distributions are not significantly different. (b) Mass chromatograms of the acetonitrile/acetic acid species (blue) and its sodium adduct (red) at 333 s. Both species coelute and have similar elution profiles. The high-intensity chromatograms are artefacts from the erroneous March samples.



## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

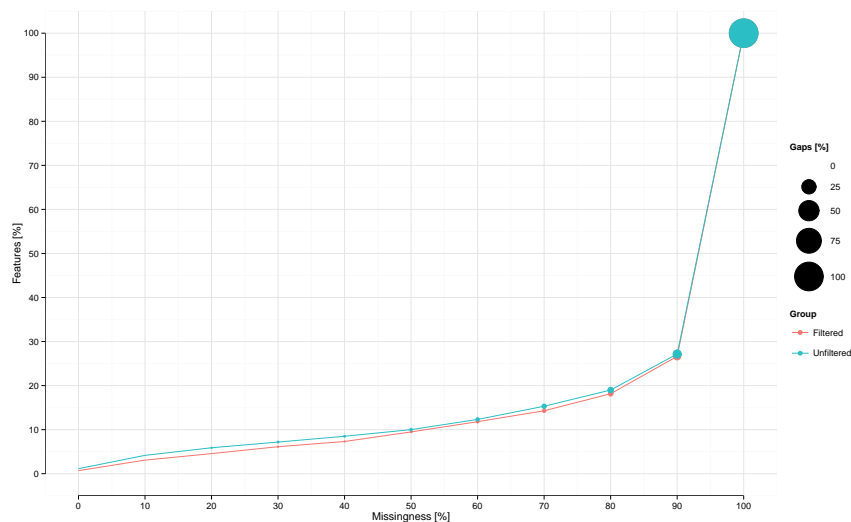
---

Name	Sum Formula	Mass Deviation	Ion Type
N-Methyl 2-pyrrolidone	C <sub>5</sub> H <sub>10</sub> NO	0.8 ppm	[M+H] <sup>+</sup>
Acetonitrile/Acetic acid	(CH <sub>3</sub> CN) <sub>n</sub> (CH <sub>3</sub> COOH) <sub>m</sub>	0.5 ppm	[A <sub>1</sub> B <sub>1</sub> +H] <sup>+</sup>
Acetonitrile/Acetic acid	(CH <sub>3</sub> CN) <sub>n</sub> (CH <sub>3</sub> COOH) <sub>m</sub>	19.0 ppm	[A <sub>1</sub> B <sub>1</sub> +Na] <sup>+</sup>
Oleamide	C <sub>18</sub> H <sub>35</sub> NO	0.5 ppm	[M+H] <sup>+</sup>
Polyethylene glycol	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	0.6 ppm	[A <sub>2</sub> B+H] <sup>+</sup>

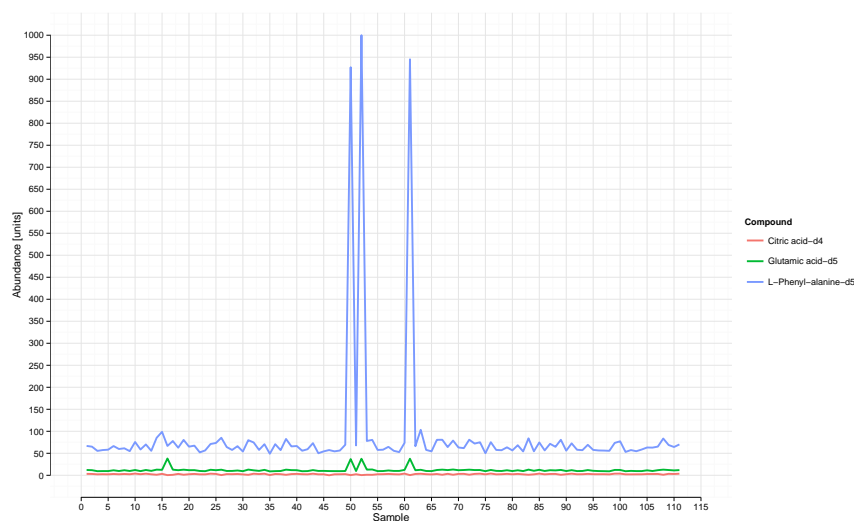
Table 2.2: Overview of the five most common interferents. The table shows the compound name, sum formula, mass deviation in ppm between the exact theoretical and found ion mass, and ion type found.

After alignment, features before 80 s and after 750 s were filtered out before isotope detection and compilation of the feature matrix for multivariate analysis. The data set was split into two. In the first data set ('filtered'), all features without isotope information were filtered out. The second data set contains all features ('unfiltered'). Isotopic envelopes serve as additional orthogonal information to improve confidence in extracted features. A data set filtered for isotopes can serve as sanity check for data analysis because features with isotopic envelopes are less likely to be irrelevant artefacts. Both data sets were used to investigate how incremental missingness affects the result matrix. Missingness is defined as percentage of absence for a feature across samples, i.e. the number of gaps. Higher missingness allows more features to be included in the matrix but may potentially distort downstream analysis. Zero missingness is ideal but impractical because of imperfections of the instrument and in data processing. The filtered and unfiltered data sets behave identical with regard to missingness (Figure 2.23a). The number of aligned features increases with allowed missingness, with a steep increase from 90% to 100%. Maximum missingness includes features that are only present in one sample, which should be considered unreliable. Stepping up to 100% missingness includes the bulk of features from unfiltered noise, hence the increase in aligned features. At 0% missingness, a total of 23 and 165 features are observed in the filtered and unfiltered data respectively. The number of gaps doubles with every 10% increase in missingness.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS



(a) Missingness bubble plot



(b) Compound intensity plot

Figure 2.23: Overview of the sample by feature missingness analysis. (a) Bubble plot of the missingness versus number of features, expressed in percent relative to the number of features at 100% missingness. The plot shows a step increase at 90% to 100%, where features only present in one sample are accepted in the matrix. At 0% missingness, 165 features can be observed in the unfiltered data. The area of the data points corresponds to the number of gaps (percent relative to the total number of gaps at 100% missingness). The number of gaps double with every 10% increment in missingness. (b) Line plot of the sample index versus intensity of Citric acid-d4, Glutamic acid-d5, and L-phenyl-alanine-d5. Intensity spikes can be observed for all March samples at index 16, 51, 53, and 52. No spike is present for L-phenyl-alanine-d5 at position 16 due to misalignment.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

Name	CID	Sum Formula	Exact Mass	[M+H] <sup>+</sup>
Citric acid-d4	16213286	C <sub>6</sub> H <sub>4</sub> D <sub>4</sub> O <sub>7</sub>	196.05211	197.05939
L-alanine-d4	12205373	C <sub>3</sub> H <sub>3</sub> D <sub>4</sub> NO <sub>2</sub>	93.07279	94.08006
Glutamic acid-d5	56845948	C <sub>5</sub> H <sub>4</sub> D <sub>5</sub> NO <sub>4</sub>	152.08454	153.09182
L-phenyl-alanine-d5	13000995	C <sub>9</sub> H <sub>6</sub> D <sub>5</sub> NO <sub>2</sub>	170.11036	171.11764

Table 2.3: Overview of the deuterated internal standards. The table shows the compound name, PubChem compound identifier (CID), molecular formula, exact mass for the deuterated compound, and exact mass for the main ion in positive ion mode.

Gaps are highly undesirable and typically need to be filled for most multivariate methods. This introduces an element of uncertainty about why the signal is absent in the sample and about the method to be used for gap filling. Gaps were filled by reverse lookup of background signals in the raw data or, in the absence of a background signal in the vicinity of the feature’s  $m/z$  value and retention time, a default value of 100,000 was set.

Given the identical behaviour of the filtered and unfiltered data sets with regard to missingness, the features in the unfiltered data set were assumed to be representative for the samples. The filtered data set was discarded. A maximum missingness of 10% enabled the detection of the internal standards (Figure 2.23b), resulting in a matrix with 680 features, and was chosen as optimum value. The standards listed in Table 2.3 were identified within a mass accuracy of 10 ppm. Spikes in abundance can be noted for the internal standards of all four March samples, similar to what is shown in Figure 2.19a, with the exception of sample 2011-03-22\_04 at position 16, where no feature was found for L-phenyl-alanine-d5 and gap filling was applied. The missing feature was incorrectly aligned and was skipped by the semi-automated analysis process. L-alanine-d4 was removed from the analysis because of its retention time at 85 s, which is borderline to the time filter at 80 s. The compound was only detected in 60% of all samples in this inconsistent region.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

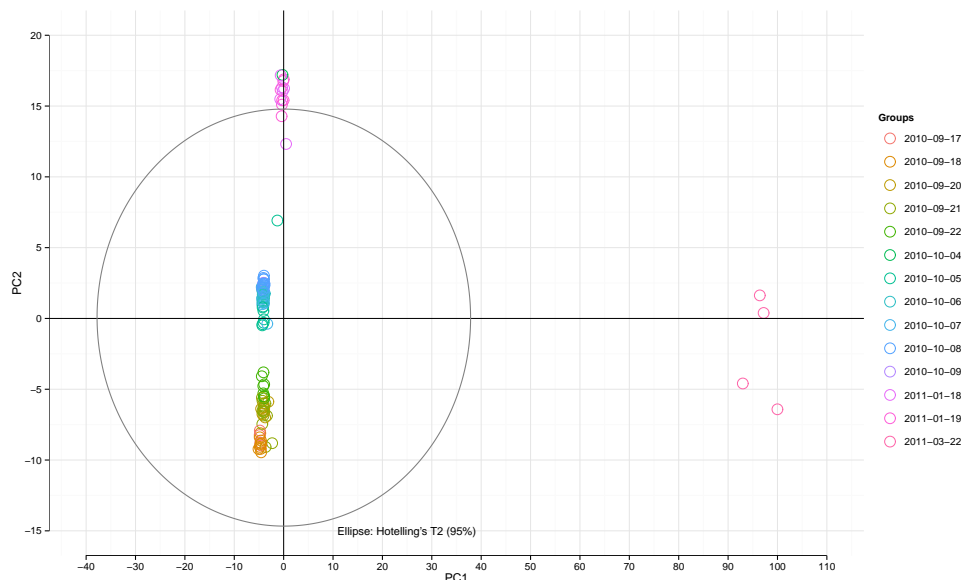
---

The resulting feature matrix was used for principal component analysis (PCA). The matrix containing 111 samples and 680 features was autoscaled before it was subjected to PCA. The result is shown in Figure 2.24a, where the first two components explain 60% of the variance (Figure 2.24b). The four March samples are distinct outliers and distort the analysis because of their high feature intensities. The samples were removed before re-analysis. The corrected matrix is shown in Figure 2.25a. The explained variance is still 60% but with greater contribution from the second principal component (Figure 2.25b). The PCA reveals distinct clustering of samples by date. Ideally, all samples should be indistinguishable because of the lack of biological variation. The differences could result from sample preparation, data acquisition, or data processing. Given that the whole data set was batch processed and the variation is not random, invalid data processing can be ruled out. The strong clustering by sample date indicates instrument-introduced variation under the assumption that the minor errors during sample preparation were random. The observed inter-aliquot variation disappears when biologically different samples are introduced (see section 3.5, Figure 3.4).

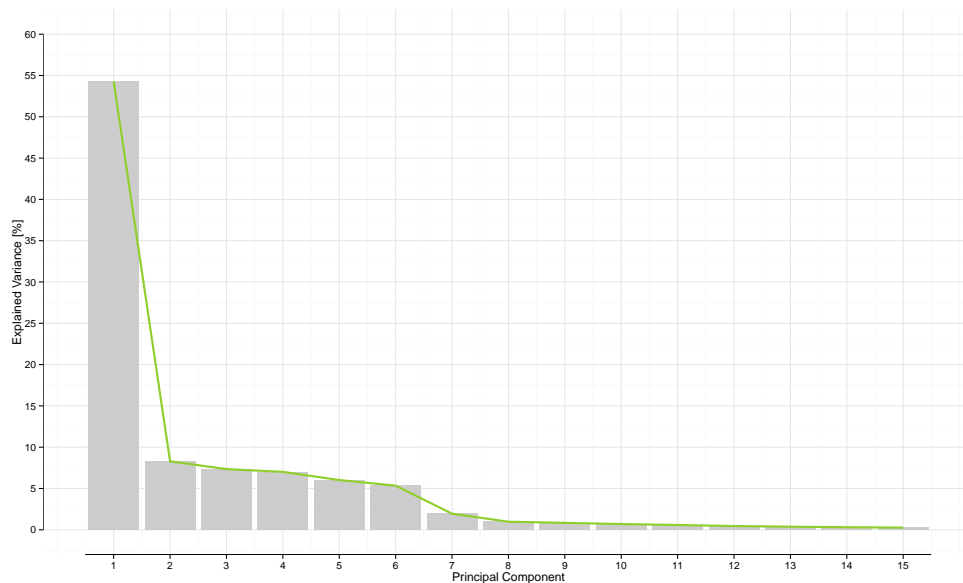
A drift can be noted for five samples acquired on 2010-09-21. Isolation of that sample group and analysis of the intensity distributions, reveals that samples acquired earlier on the day (index 2, 3, 9, 15, and 20) have shifted distributions towards higher intensities (Figure 2.26). Even though the shift is not significant, it is picked up in this analysis due to the absence of big variation, e.g., biological variation. This shift is responsible for the drift observed in the PCA.

In this case study, MassCascade was used to process LC-MS<sup>n</sup> data of 113 quality control samples from wild-type *Solanum lycopersicum*. Because of the lack of biological variation, variation introduced by sample preparation and instrumentation could be picked up. The clustering seen in PCA are non-random and could be explained, ruling out significant errors through the actual data processing. This demonstrates that MassCascade can be used for fingerprinting experiments and data exploration.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS



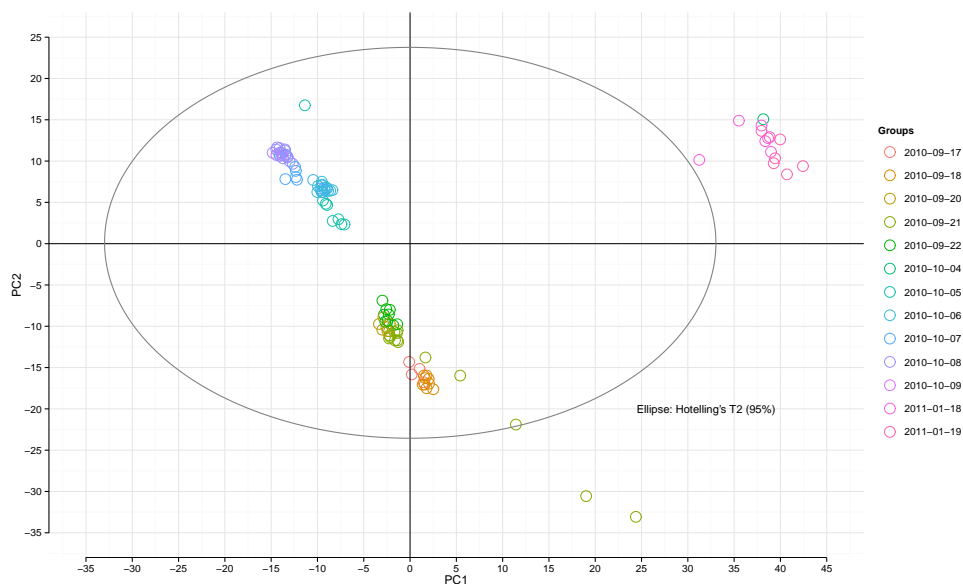
(a) Principal component analysis



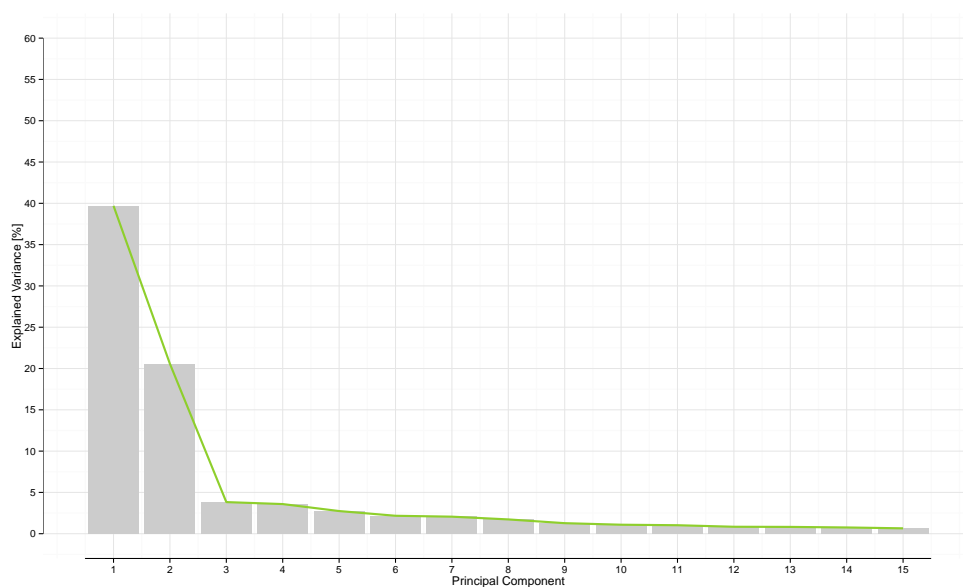
(b) Variance explained by PCs

Figure 2.24: (a) Principal components analysis of the aligned features of all samples. The samples are coloured by acquisition date. A confidence region is shown as ellipse using Hotelling's  $T^2$  statistic. The March samples are strong outliers, obfuscating the plot. (b) Bar chart of the explained variance by principal components in percent. The first two principal components explain 60% of the overall variation.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS



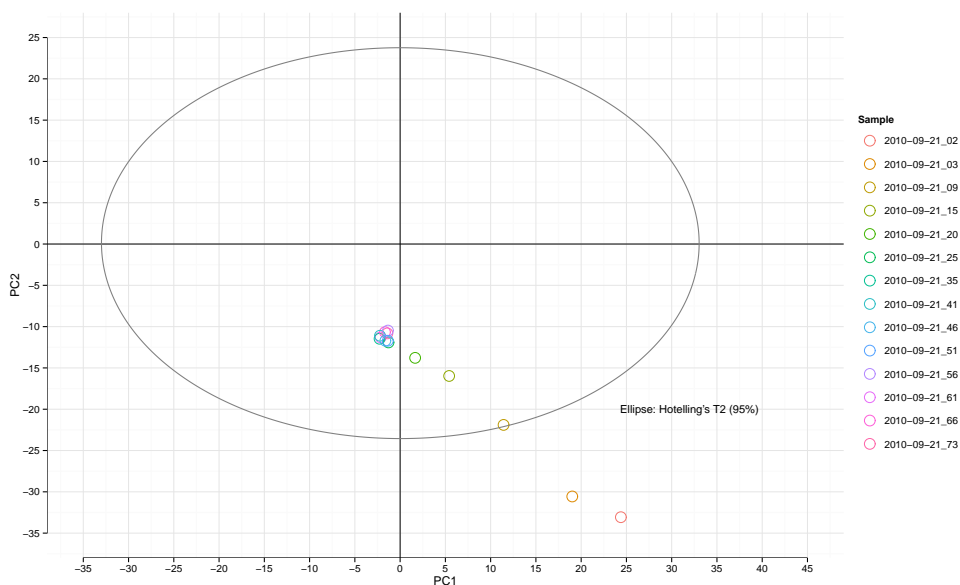
(a) Principal component analysis



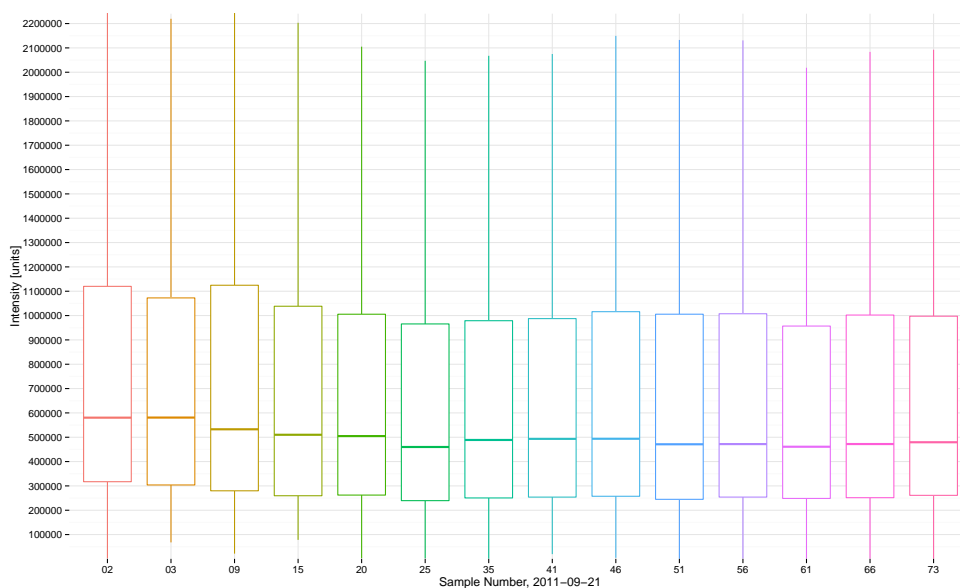
(b) Variance explained by PCs

Figure 2.25: (a) Principal components analysis of the filtered features of all samples. The samples are coloured by acquisition date. A confidence region is shown as ellipse using Hotelling's  $T^2$  statistic. The samples cluster by acquisition date. Five samples of the group from 2010-09-21 exhibit a drift away from the main cluster. (b) Bar chart of the explained variance by principal components in percent. The first two principal components explain 60% of the overall variation.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS



(a) Principal component analysis



(b) Intensity boxplots

Figure 2.26: (a) Principal components analysis of the samples from 2010-09-21. The samples are coloured by run index. A confidence region is shown as ellipse using Hotelling's  $T^2$  statistic. Five samples drift away from the main cluster. (b) Corresponding boxplot of the intensity distributions of the samples. The same colour code is used. Samples at index 2, 3, 9, 15, and 20 show a marginal shift in distribution to higher intensities. This shift maps to the drift observed in the principal component analysis.

### 2.5.4 Performance and Scaling of the Core Library

In contrast to the core library, the KNIME plug-in is tightly bound to the KNIME framework and performance varies greatly from architecture to architecture and available resources. The plug-in is configured to serialize all data to disk, ensuring scalability within the workflow environment as long as sufficient memory is given to perform single sample operations. Execution speed is primarily limited by data read-write cycles. Other factors are negligible. Consequently, only the core library was systematically benchmarked. The desktop specifications and run time for the study employed in this chapter are listed at the end as rough guide only.

The core library can be run in file or memory mode. In file mode, execution speed is limited by data read-write cycles like in the plug-in. It is assumed that the plug-in is used for data local processing. Hence, the performance of the MassCascade core library was tested in a server environment with a varying number of threads from 2 to 20 and 16 GB of memory. Every thread deals with one sample. The data processing steps of the fingerprinting case study were reproduced for this test.

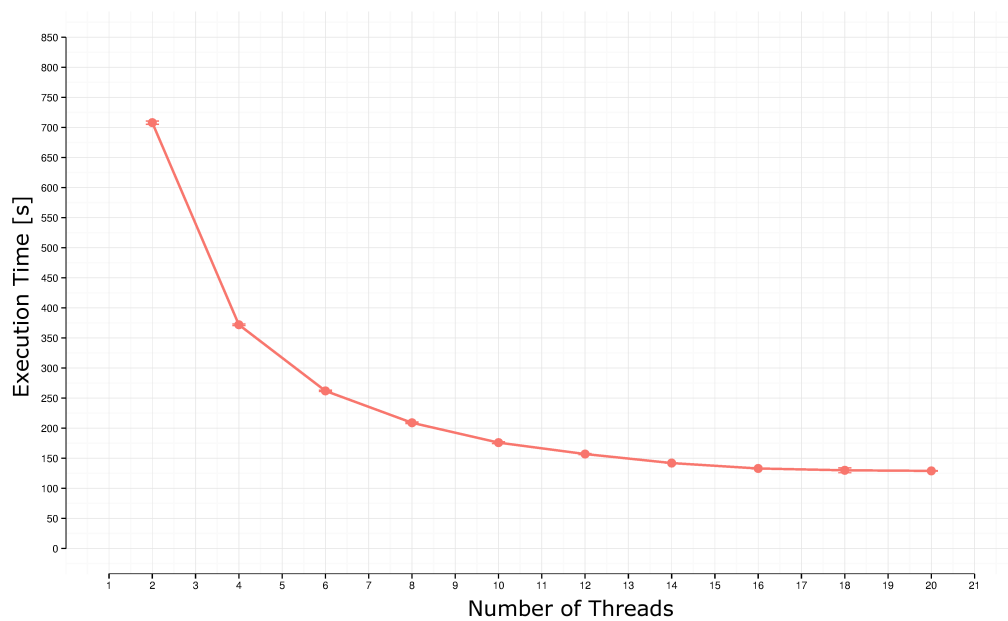
The execution time for the program decreases with the number of allocated threads following a power law. The log-log plot defines the relationship further as monomial,  $y = ax^k$  (Figure 2.27). MassCascade uses threads as atomic unit for sample processing. Thus, it scales well in server environments that deal efficiently with threading. Given that the total execution time does not decrease linearly with an increasing number of threads, it can be assumed that MassCascade is primarily I/O or memory bound. That is, the memory accessing costs for frequent loading of non-aligned data is the limiting factor.

The evaluation (section 2.5) was run in KNIME using the MassCascade plug-in on a desktop, core i7 740Q @ 1.73 GHZ, with 3 GB memory allocated to the KNIME environment. The plug-in took 8 hours to run the complete workflow for the centroided dataset (total file size: 4.6 GB).

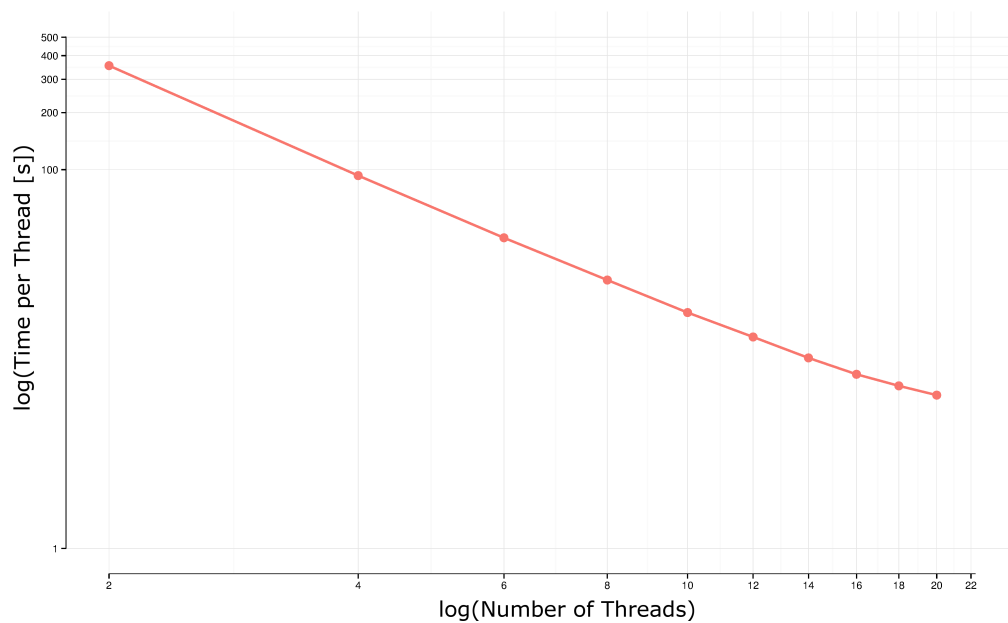


## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---



(a) Performance line plot



(b) Log-log plot

Figure 2.27: (a) Line plot of the thread number versus the total execution time. An increasing number of threads reduces the execution time following a power law. (b) Log-log plot of (a). The depicted linear relationship can be described as monomial.

## 2.6 Technical Validation

A small well described case-control study is used to assess the reliability and performance of the tool after its use had been demonstrated in the analysis of tomato samples (section 2.5).

Technical assessment is typically carried out against well-described biological data sets with known outcome<sup>[107,128,147,158,163]</sup>, mixture samples<sup>[155]</sup>, synthetic data sets<sup>[101,104]</sup>, or combinations of these<sup>[156,231]</sup>. Here, a data set from a dilution series of *Arabidopsis thaliana* seed and leaf extracts is used. The data set is freely available as part of the publication by Tautenhahn *et al.*<sup>[231]</sup>.

The performance of MassCascade is evaluated by its ability to isolate relevant features from the data set using its data processing methods. Following the approach by Tautenhahn *et al.*, the problem to isolate relevant features is treated as information retrieval problem and can thus be evaluated by values for precision and recall (sensitivity). The *centWave* method for feature isolation – discussed in the aforementioned publication – is used as a reference point.

Precision  $P$ , the ratio of the number of real features  $TP$  to the total number of features  $N$ , and recall  $R$ , the fraction of the number of real features  $TP$  to the total number of real features  $NP$ , are combined into a F-score  $F$  for interpretability. A F-score of 100% equals perfect precision and recall. False positives and false negatives reduce the F-score<sup>[239]</sup>.

$$P = \frac{TP}{N} \tag{2.18}$$

$$R = \frac{TP}{NP} \tag{2.19}$$

$$F = \frac{2 * R * P}{R + P} \tag{2.20}$$

The following section outlines the design of the technical validation, including a brief description of the data set and the compilation of a ‘ground truth’ to determine the total number of real features<sup>[231]</sup>.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

### 2.6.1 Methods

The data set consists of eight samples. Each sample is measured in ten technical replications. The dilution series contains mixtures of solvent and either seed or leaf extracts in varying proportions (solvent/seed/leaf [v/v/v]): 0/100/0, 25/75/0, 50/50/0, 75/25/0, and 0/0/100, 25/0/75, 50/0/50, 75/0/25.

The pure seed and leaf samples (0/100/0 and 0/0/100) are separately processed and the ten technical replications aligned for each case using Obiwrap with default parameters. Removing features present in less than seven out of the ten technical replicates yields the ‘ground truth’ for the subsequent analysis of the seed and leaf dilution series respectively, i.e. the total number of real features.

Because MassCascade is compared to XCMS’s *centWave* method, the ‘ground truth’ is also estimated using XCMS *centWave*. The intersection of both ‘ground truths’, i.e. matching features within a 75 ppm  $m/z$  tolerance window, is used as the combined ‘ground truth’ for the following analysis. The tolerance window as well as the parameters for *centWave* are derived from the original publication.

Subsequently, the diluted samples are processed by each tool separately and evaluated against the combined ‘ground truth’. The samples for the 25%, 50%, and 75% dilutions of the seed and leaf samples are processed identically to the pure samples – noise reduction, Durbin Watson filtering, feature set generation – with the exception of feature alignment. Isolated features are then matched to the total number of real features. Matching features are considered true positives, all other features false positives. Feature alignment is not required because the pre-processing and processing algorithms are evaluated in this scenario, which work on a sample-by-sample basis to isolate relevant features.

The 80 samples (the stock solution plus three dilutions per extract with ten replications each) are converted from *mzXML* format to *mzML* using ProteoWizard<sup>[230]</sup> before they are processed using the MassCascade KNIME plug-in. Default parameters are used throughout with intensity thresholds lowered to 10 units.

## 2.6.2 Results & Discussion

The combined ‘ground truth’ for the seed and leaf extract contained 1331 and 619 features respectively. The *centWave* method had picked up 1.5 times more features than MassCascade after filtering. The *centWave* method was developed for peak picking and no further filtering had been applied, e.g. based on a Durbin Watson statistic.

It should be noted, that the *centWave* method was run with the recommended parameters from the publication by Tautenhahn *et al.*<sup>[231]</sup>. The algorithm has been modified and improved since 2008 and the F-score values obtained in this analysis should be understood as reference points only.

Figure 2.28 summarises the results. The MassCascade pipeline performs equally well to the *centWave* method in the case of the leaf extracts and is less successful by approximately 10% in the case of the seed extracts.

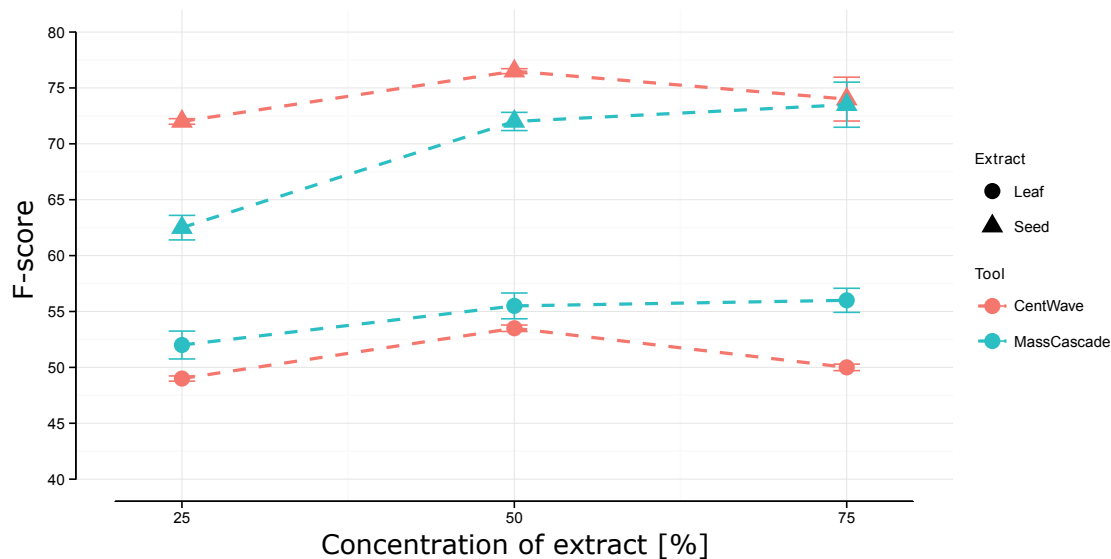


Figure 2.28: Line plot of the F-scores for the informational retrieval challenge. Mass-Cascade (blue) is compared to *centWave* (red) as reference. The leaf and seed scores are shown as circles and triangles respectively, including their standard errors. The samples of the dilution series are made up of 25%, 50%, and 75% leaf or seed extract. In this scenario, MassCascade performs comparable to the *centWave* method.

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

The decrease in *centWave*'s F-score for the 75% samples is an artefact and should be evaluated with caution: the 'ground truth' is larger for the *centWave* method. The MassCascade pipeline is more cautious in this scenario because it filters the feature set and is thus the limiting factor with regard to the size of the combined 'ground truth'. Because *centWave* picked up significantly more low-intensity features for the 75% samples (high  $N$ ) but the combined 'ground truth' ( $NP$ ) was limited in numbers by MassCascade, its F-score decreased.

MassCascade appears to be more conservative with regard to feature selection, i.e. it restricts the number of final features returned downstream of the analysis pipeline, but shows good precision and recall for the those features as reflected in the F-score of the dilution series.

### 2.7 Conclusion

MassCascade offers a modular, step-by-step solution for processing mass spectrometry data. The plug-in enables the prototyping of complex data analysis workflows to explore data sets. Intermediate results can be inspected after every processing step. The plug-in's ease-of-use and the ready availability of additional nodes with complementary methods for further data analysis are its key features. It works reliably for the isolation of relevant features and can help in the standardisation and structuring of data processing and analysis, much needed in the field of metabolomics, where the diversity of instruments and variables can be challenging. Outstanding work involves the need to fine-tune the methods for different types of mass spectrometers to take into account differences between mid- and high-resolution spectrometers and different chromatographic methods.

### 2.8 Software Availability

MassCascade and its plug-in MassCascade-KNIME have been released under the GNU General Public License version 3. External tools such as the alignment

## 2. INFORMATICS FOR LC-MS<sup>n</sup> ANALYSIS

---

program Obiwarrior, need to be installed separately and are not included in the release of MassCascade or MassCascade-KNIME.

### 2.8.1 Update Site

The projects are deposited on the code hosting website BitBucket. Supporting information is provided in the projects' *wiki* pages. An update site for the KNIME plug-in installer is provided with the latest version. The entry pages are listed below:

- MassCascade:  
<https://bitbucket.org/sbeisken/masscascade/>
- MassCascade-KNIME:  
<https://bitbucket.org/sbeisken/masscascadeknime/>
- KNIME Update Site:  
<https://bitbucket.org/sbeisken/masscascadeknime/wiki/release>

### 2.8.2 Extensions

Extensions for the core library or KNIME plug-in can be added following the help pages provided on the BitBucket's project page.

### 2.8.3 Example Workflows

Example data can be retrieved from the MetaboLights database. The quality control data set used in section 2.5 was obtained from study "MTBLS36: Metabolomic Study of different Cultivars of Tomatoes". Several example workflows illustrating different aspects of the plug-in are deposited on the project's wiki page.

The work has been published: Beisken *et al.*: MassCascade: Visual programming for LC-MS data processing in metabolomics. *Molecular Informatics 2014*

---

# KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

## 3.1 Introduction

An identification framework was implemented in MassCascade and tested on a study about the ripening behaviour of tomato cultivars. Identification in MS-based metabolomics has been introduced in chapter 1.1.6 of the introduction. General problems with metabolite identification have been pointed out such as false positives and lack of stereo-information. Data-driven techniques were explained – simple  $m/z$  lookups, spectra queries – and the idea of additional orthogonal information was introduced combined with confidence rankings for multiple annotations.

In contrast to the challenge of identifying *known* metabolites, i.e. previously encountered metabolites that are available in a metabolite database, *unknown* metabolites have – by definition – not been previously encountered. *Unknown unknowns* are novel compounds that have not been previously described; they are not contained in compound or spectral databases. *Known unknowns* refer to chemical structures that are stored in generic compound databases but have

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

not been identified in metabolomics experiments or linked to specific metabolite databases. These metabolites cannot be retrieved via the process described above. Instead, sum formulas need to be calculated for unknown features based on their  $m/z$  values. Mathematical and chemical rules can be applied to constrain the number of resulting sum formulas<sup>[240]</sup> in combination with high mass accuracy (ma < 1 ppm), but it has been shown that even sub-ppm mass accuracy does not necessarily result in a unique sum formula for a given mass<sup>[241]</sup>. The calculated sum formulas represent large numbers of constitutional isomers. Isotope and fragmentation information can be used to narrow down the constitutional isomer space to the point where manual curation and interpretation becomes feasible. Tools for calculation of sum formulas have been developed to automatize this process<sup>[242]</sup>.

For the identification of unknown metabolites, tandem mass spectrometry has proven most valuable<sup>[243,244]</sup>. Either through *in silico* fragmentation of isomers and subsequent matching to measured fragmentation spectra<sup>[245]</sup> or *a priori* sum formula calculation of fragments and subsequent re-assembly of the parent molecule across multiple MS<sup>n</sup> level under the constraints provided by the fragment sum formulas<sup>[246]</sup>. With advances in MS technology, fragmentation approaches appear to be most promising for untargeted metabolite identification<sup>[247-250]</sup>. The identification of true *unknown unknowns* remains a specialist application<sup>[87,242]</sup>. At the moment, the challenge is to fast and unambiguously identify signals from mass spectrometry data in metabolomics without time-consuming manual intervention. This is believed to be best achievable through identification frameworks that aggregate information for information-guided decision making<sup>[17,38]</sup>.

The aim of this chapter is to evaluate the implemented scoring framework and to identify metabolites involved in the ripening of tomato. After initial processing and validation of a metabolomics study about tomato ripening, features believed to be involved in ripening are singled out for identification with the scoring framework. Study validation is carried out using principal component analysis (PCA) to investigate batch effects and sample outliers. Features for identification are determined through a feature orientated approach based on orthogonal partial least squares (OPLS) rather than through a metabolite orientated approach. The



### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

selected features are run through three different scenarios to score annotations retrieved from different sources, simulating the cases of *known*, *known unknown*, and *unknown unknown* metabolites. The scenarios and subsequent interpretation of the highest-scoring metabolite annotations are used to evaluate the scoring framework.

In addition, the study’s data sets and findings are compiled as Datument, i.e. a “hyperdocument for transmitting and preserving the complete content of a piece of scientific work”<sup>[251]</sup>, and shared publicly including all information, lending itself to applications in metabolic cross-study comparisons or investigations into reproducibility of data analysis in metabolomics.

Finally, technical validation is carried out using an open data set for benchmarking to evaluate the implemented scoring framework using solely MS data without the context of a biological study.

#### 3.1.1 Open Data

Data cannot be valued without knowledge<sup>[252]</sup>. Metabolite identification is most beneficial to the community if all information and raw data from the study is publicly available. Sharing the totality of collected information also increases the value of the data itself. Use of established standards, e.g. well accepted recommendations by the HUPO Proteomics Standards Initiative for reporting or accepted formats such as IUPAC’s InChI line notation for small molecule management, and semantic publishing enables re-use of data and may increase study validity<sup>[253]</sup>.

Datuments, the term was originally coined by Murray-Rust *et al.*, have been described as scientific enablers that are required by complex data sets<sup>[254]</sup>. Compiling Datuments and sharing data including the underlying semantic models is an interesting challenge<sup>[255]</sup> that is met by recent initiatives such as a new content type called the ‘Data Descriptor’ by the Nature Publishing Group, which focusses primarily on data. In metabolomics, the challenge, *inter alia*, is met by the MetaboLights database<sup>[143]</sup>. Together, both resources enable the scien-

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

tific community to capture standardized Datuments in an open access and open source way that is semantically accessible.

As part of the work on knowledge-based compound identification, data and techniques used have been standardized and reported using purely open access and open source solutions while attempting to preserve as much information as possible, following current efforts in the metabolomics community on standardized data<sup>[141]</sup>.

## 3.2 Metabolite Identification

This work assumes certainty of chromatographic selectivity<sup>[256]</sup> on the premise that compounds have either been separated during chromatography or that co-eluting molecular compounds have been resolved using peak picking and deconvolution. The identification process can be broken up into several complementary, discrete steps<sup>[257]</sup>. Many studies focus only on the one or two most relevant of these steps, e.g. fragmentation trees<sup>[87,242,258]</sup>. Some studies take additional factors into account as proof-of-principle, e.g. retention time prediction using artificial neural networks<sup>[259]</sup>. The following section outlines a semi-automated approach consuming several factors for initial identification and subsequent rationalisation and ranking.

### 3.2.1 Identification Factors

The workflow for known and unknown metabolite identification is shown in Figure 3.1. The identification process starts from annotated feature sets, where annotations are retrieved from either compound or spectra databases. Feature sets can be sample- or consensus-based before initial annotation: *Consensus Feature Sets* can be generated in a ratio analysis-like fashion, coined RAMSY by Gu *et al.*<sup>[260]</sup>. For a group of samples, all feature sets are aligned cross-sample and individual features matched within a given  $m/z$  tolerance (in ppm). Contrary to other methods, equidistant binning has been replaced by adaptive intelligent

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

binning<sup>[261]</sup> to maximise feature matching cross-sample. Because of small time shifts, equidistant binning is more likely to miss individual features in a small subset of feature sets. The mean and standard deviation of individually aligned feature intensity ratios (equation 3.1 with  $I_{main}$  being the most intense feature), form the basis of the resulting consensus feature set: the division of the average feature’s intensity ratio ( $\bar{I}_r$ ) by the intensity ratio distribution’s standard deviation gives the new intensity ( $I'_i$ , equation 3.2) of the feature. Consensus feature sets that are generated using this approach highlight features that are consistent across samples. In turn variable features with large standard deviations are de-emphasised in a spectrum match.

$$\bar{I}_r = \frac{1}{n} \sum_{i=0}^n \left( \frac{I_i}{I_{main}} \right) \quad (3.1)$$

$$I'_i = \frac{\bar{I}_r}{\sigma_{I_r}} \quad (3.2)$$

The identification factors are implemented as optional filters that can be turned on or off. Filters act on pre-perceived information from data processing steps, e.g. detected isotope annotations. The filters remove identity annotations, i.e. molecule annotations from queries against spectra or compound libraries, or alter the identity annotations’ overall score based on the presence or absence of the filter criteria (see the following section for details on the scoring scheme used).

- Feature sets are used as initial guess about correlated features in MS<sup>1</sup><sup>[127]</sup>. These serve as basis for information inferred from neutral losses<sup>[262]</sup> and isotopes (see chapter 1.1.4).
- If no identity annotation is provided for a feature that has fragmentation spectra, the feature is regarded a known unknown and the PubChem database is queried for possible annotations based on the feature’s  $m/z$  value.
- Molecule annotations are filtered by their elemental composition, including the common organic subset CHNOPS + Halogens.
- The isotope filter selectively removes annotations whose theoretical isotopic

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

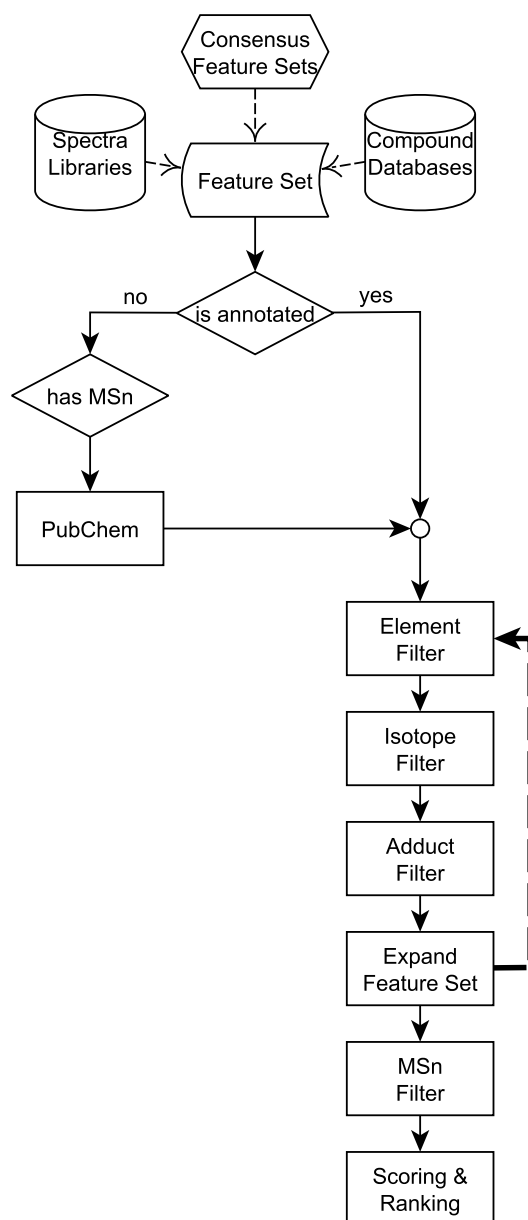


Figure 3.1: Workflow for known metabolite and known unknown metabolite identification. (Consensus) feature sets are annotated through spectra libraries or compound databases before each identity annotation is filtered for the organic element subset CHNOPS + Halogens, isotope presence and abundance matching, and adduct support, before feature sets are expanded for multiply annotated features in the same feature set. If no initial identity annotations are present for a feature with fragmentation information, the compound database PubChem is queried for possible annotations. The MSn filter matches the explained fragmentation features against all fragmentation features before the total scoring and ranking is reported. The overall score is the result of contributions from each individual filter.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

envelope does not correspond to the measured intensity ratios. The filter selectively checks features with perceived isotope annotations. Theoretical and computational intensity ratios  $\left(\frac{I_{M+X}}{I_M}\right)$  have to match within a 10% tolerance window.

- The adduct filter selectively checks the presence of perceived adduct annotations and lowers the ranking of annotations that do not feature adducts within a feature set.
- The MS<sup>n</sup> filter is based on the MS<sup>n</sup> enumeration function. Enumeration of fragmentation spectra refers to brute-force feature fragmentation applied on identity annotated features in MS<sup>1</sup> that have fragmentation spectra. Molecular structures are deterministically fragmented in a breadth-first fashion down to a pre-defined depth of or minimum molecular weight. This approach is similar to current top-down standards in LC-MS<sup>n</sup> metabolomics. Noteworthy, a more systematic approach based on quantum mechanics was proposed by Galezowska *et al.*<sup>[244]</sup>. The MS<sup>n</sup> filter matches the spectral vector consisting of the subset of explained, i.e. enumerated, fragmentation features against the complete spectral fragmentation vector and adjusts the ranking based on that score.
- The missingness filter works not on the annotation level but directly on the feature set level. It removes features that cannot be found across more than x% of samples in its designated group. For sample-by-sample reporting, rather than batch reporting, the missingness filter can trim identifications with respect to global metabolite identifications across the batch or sample group. It shares properties with the approach used for the compilation of consensus feature sets and should therefore not be used in combination with these.

This purely annotation or information driven approach results in a ranked list of metabolite identifications. It has the advantage of ease-of-use due to a lack of direct dependencies beyond spectral and compound databases, which are required for any form of known or known unknown metabolite identification. If multiple features have metabolite identifications in a single feature set, the feature set

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

is expanded, i.e. the set is copied and reported individually together with each metabolite identification. Additional identification factors could include models based on characteristic spectral fingerprints of individual molecules derived from large study-specific training sets<sup>[263]</sup>.

#### 3.2.2 Scoring Schemes

The developed scoring scheme is based on the filters for the identification factors. It starts from an initial score of zero, ignoring the obtained score from the identity annotation process: the normalized score of a simple  $m/z$  lookup or spectrum match that relates either to the  $m/z$  deviation between target and library feature or the query score returned by the web database. This ‘two-pass’ system – independent annotation and rationalization – avoids potential biases introduced by spectrum matching.

For local library queries, a robust weighted scoring scheme ( $MF$ ) has been implemented consisting of the dot product ( $F_D$ ) of the weighted target ( $T$ ) and library ( $L$ ) spectral vectors (with elements  $z$ ) and their ratio of feature pairs ( $F_R$ ). The elements of a spectral vector,  $z$ , are weighted by their  $m/z$  and intensity value and sorted in ascending order by their respective  $m/z$  values.  $n$  is the number of features in either vector or their intersection ( $n_{L \cap T}$ )<sup>[264,265]</sup>.

$$MF = \frac{n_T F_D + n_{L \cap T} F_R}{n_T + n_{L \cap T}} \quad \{MF \in \mathfrak{R} \mid 0 \leq MF \leq 1\} \quad (3.3a)$$

$$F_D = \frac{(\sum z_L z_T)^2}{\sum z_L^2 \sum z_T^2} \quad \{F_D \in \mathfrak{R} \mid 0 \leq F_D \leq 1\} \quad (3.3b)$$

$$F_R = \frac{1}{n_{L \cap T}} \sum_{i=2}^{n_{L \cap T}} \left( \frac{z_{L,i} z_{T,i-1}}{z_{L,i-1} z_{T,i}} \right) \quad \{F_R \in \mathfrak{R} \mid 0 \leq F_R \leq 1\} \quad (3.3c)$$

$$z = I^p m z^q \quad (3.3d)$$

The local search score is normalized between 0 and 1,000. The intensity and  $m/z$  weights are set to  $p = 0.6$  and  $q = 1.5$  according to recommendations by Stein *et al.*<sup>[264,265]</sup>, giving greater emphasise to higher  $m/z$  values. Adding the term  $F_R$  to

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

Filter	Score	Evidence
Element	{0}	Weak
Isotope	$\{0 \leq x \leq 200\}$	Medium
Fragment	$\{0 \leq x \leq 500\}$	Medium, Strong
Relation	$\{0 \leq x \leq 100\}$	Weak
Missingness	$\{0 \leq x \leq 200\}$	Weak

Table 3.1: Summary of the ranking score scheme. Five filters are implemented that contribute between 0 and 500 points to the total score of 1000. The value of the score depends on the quality with which a metabolite annotation passes a filter. The evidence column shows a three category scheme to easily cluster ranked results.

the dot product  $F_D$  gives additional weight to the intensities of matching signals, improving the power of the score when  $n_{L \cap T}$  is very large. Alternative scores have been reported that provide a statistical approach to spectrum matching taking global spectra similarity and molecular properties into account<sup>[266–268]</sup>. These approaches could form highly beneficial extensions when more comprehensive metabolomics spectral libraries have been established.

The weighting scheme subsequently used by the filters is listed in Table 3.1. Each filter adds to the overall score of 1,000. Fragment filtering yields the highest contributions because of its importance in identification, e.g. in contrast to adduct information (relation), which only adds confidence. The isotope and fragment filters have intrinsic thresholds: annotations that do not match the isotope filter and annotations which fragment scores fall below 100 (a spectral match of less than 10%) are automatically removed.

### 3.3 Materials and Methods

Two data sets were used for metabolite identification, both provided by the Syngenta AG. The data sets were co-produced as part of a larger study on Tomato ripening, which is about to be submitted by the Syngenta AG. The high-resolution data sets used here were analysed but ultimately not included in

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

the final manuscript. The metadata of the two studies was collected and compiled for submission to MetaboLights<sup>[23]</sup>. The study has been submitted to Nature Scientific Data: Beisken *et al*: Metabolic differences in ripening of wild and mutant cultivars of *Solanum lycopersicum*. The following outlines the purpose and design of the studies.

The ripening behaviour of *Solanum lycopersicum* and three ripening-inhibited mutants was explored using a LC-MS<sup>2</sup> assay. The plant material was harvested in five to ten day intervals up to the point of flowering and daily after. Each genotype was grown in triplicate. Each block of 52 plant samples was measured using a group-randomization setup with interspersed quality controls: blank injections (solvent), pooled samples (mix, see table below), and aliquots of a pooled in-house standard tomato reference. Randomization groups were defined by day of harvest.

The study resulted in a rich metabolic data set about the ripening behaviour of wild type and mutant tomato. Quality measures were taken into account during study design facilitating data analysis and enabling filtering of unintended biological variation in the data. A total of 58 distinct reference standards were measured on the same instrumental setup to aid metabolite identification and consequently model validation via biological interpretation. The reference compounds are those used by Syngenta for their in-house projects.

#### 3.3.1 Tomato Cultivars

Wild type *Solanum lycopersicum* (Ailsa Craig, AC<sup>++</sup>) and three ripening inhibited AC<sup>++</sup> mutants were used in this study (Table 3.2): non-ripening (NOR), ripening-inhibited (RIN), and colourless non-ripe (CNR) tomatoes<sup>[269–271]</sup>. The plants were grown in 24 cm-diameter pots in M3 compost (Levington Horticulture, Ipswich, and Suffolk, UK) and watered daily under standard greenhouse conditions. Flowers were sampled in five to ten day intervals up to anthesis (10, 15, 20, 30, 40) and daily post-anthesis (47, 48, 49, 50, 51, 52, 53, 54). Breaker fruit were defined as those showing the first signs of ripening-associated colour change from green to orange. Non-ripe mutants were taken at day 49 as post-anthesis



### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

Name	Category	Description
AC <sup>++</sup>	wild type (WT)	Alisa Craig variety
NOR	monogenic mutant	as WT except for non-ripening locus
RIN	monogenic mutant	ripening inhibited mutant
CNR	monogenic mutant	colourless non-ripe mutant
Mixed	n/a	pooled AC <sup>++</sup> , NOR, RIN, CNR, TomQC sample
TomQC	n/a	standard in-house tomato aliquots
Blank	n/a	blank sample with solvent

---

Table 3.2: Overview of sample types used in the study of different tomato cultivars. The study design included four tomato cultivars: AC<sup>++</sup>, NOR, RIN, and CNR, interspersed with pooled samples, quality control aliquots from an in-house standard tomato, and blanks consisting of solvent only.

equivalents to breaker WT fruits. All plant samples were taken at the same time each day, frozen in liquid nitrogen, and stored at  $-70^{\circ}\text{C}$  until required.

#### 3.3.2 Sample preparation

Stock standard solutions were prepared for the analytical reference standards at a concentration of  $1000\ \mu\text{g}/\text{mL}$  in 20/80 HPLC analytical grade Ethanol/Water and then diluted 10x for injection.

Tomato samples were subjected to an untargeted metabolite analysis by LC-MS/MS of polar extracts. Approximately 30 mg of dried tomato tissue was extracted with HPLC analytical grade Ethanol/Water 20:80. The polar extracts were diluted 10:1 with water and injected underivatized. Samples were acquired in three batches on 17, 20, and 21 September 2010 in positive ion mode and on 23, 24, and 27 September 2010 in negative ion mode.

#### 3.3.3 Chromatography

All 219 samples were run on a Waters Acquity<sup>TM</sup> UPLC system (Waters Corporation, USA), HSS T3 150 x 2 mm,  $1.7\ \mu\text{m}$  particles, UPLC column at  $30^{\circ}\text{C}$

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

oven temperature. Each batch consisted of 73 samples with multiple quality control samples. The solvents used for the assay consisted of 0.2% Formic Acid (Solvent A) and 98/2/0.2 Acetonitrile/Water/Formic Acid (Solvent B). Gradient [time (min) / %B] starting at flow rate 0.25 mL/min: 2.5/0, 7.5/10 (flow rate to 0.4 mL/min), 10.0/100, 12.0/100, 18.0/0, 25.0/0. Aliquots of 2  $\mu$ L were injected.

#### 3.3.4 Mass Spectrometry

The compounds were detected using a Thermo LTQ Velos Orbitrap mass spectrometer operating in positive and negative Electrospray ionization (ESI) mode at a resolution of 30,000 with a scan range from 85-900  $m/z$  and 95-900  $m/z$  respectively. MS/MS spectra were obtained in a data dependent manner through higher-energy collisional dissociation (HCD, normalized collision energy: 50.0) at a resolution of 7,500: The two most intense mass spectral peaks detected in each scan were fragmented to give MS2 spectra (100-900  $m/z$ ). Full scan data was acquired in FT (accurate mass) mode, MS/MS spectra were acquired in centroid mode. The LTQ Velos Orbitrap used the Xcalibur control software version 2.1.0 for data acquisition. Reference standards were acquired using the same protocol and experimental setting.

#### 3.3.5 Reference Standards

Reference standards were commercially purchased from Fluka Analytical, Sigma-Aldrich, and C/D/N Isotopes or prepared in-house. Pooled in-house standard tomato reference was prepared from the shop bought Angelle variety: mashed up in bulk and aliquoted out following the protocol outlined above.

#### 3.3.6 Data Deposition

All samples used in this study have been submitted to MetaboLights at the European Bioinformatics Institute (EMBL-EBI). Each MetaboLights entry contains

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

protocols about sample collection, extraction, chromatography, mass spectrometry, metabolite identification, and data transformation. The study was metadata tagged using the Investigation/Study/Assay (ISA) suite<sup>[272]</sup> in Isatab format a tab-separated text files for experimental information. In addition, metabolites identified and/or annotated were stored as mzTab<sup>[273]</sup> compatible tab separated file provided as MetaboLights’s IsaCreator plug-in extension.

#### Data Record: Tomato Study Samples

MetaboLights’s accession MTBLS36 contains the study samples: 442 LC-MS<sup>2</sup> files (*.mzML*, 64-bit) acquired in continuous mode: 219 in positive and 223 in negative ion mode. Consistent file names are composed of <acquisitionDate>\_<runId>\_<sample>\_<sampleTime>.

#### Data Record: Reference Standards

MetaboLights’s accession MTBLS38 contains the reference standards: 71 LC-MS<sup>2</sup> files (*.mzML*, 64-bit) acquired in continuous mode: 43 in positive and 28 in negative ion mode. Chemical names are used as file names and linked to the ChEBI database<sup>[274]</sup>.

## 3.4 Data Processing and Transformation

Processing of raw data was identical for both data sets up to cross-sample feature alignment. Non-targeted LC-MS datafiles were converted to *mzML* format using the program ProteoWizard<sup>[230]</sup>. Vendor-based peak picking was enabled for MS<sup>1</sup>. The resulting *mzML* files were processed with MassCascade in KNIME (see Figure 3.2 for an overview of the core workflow).

The data were time and  $m/z$  filtered before noise was reduced and features extracted with 10 ppm mass accuracy. A Durbin-Watson filter was applied with a lenient threshold of  $DW = 2.8$ . Random signals are estimated to be around

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

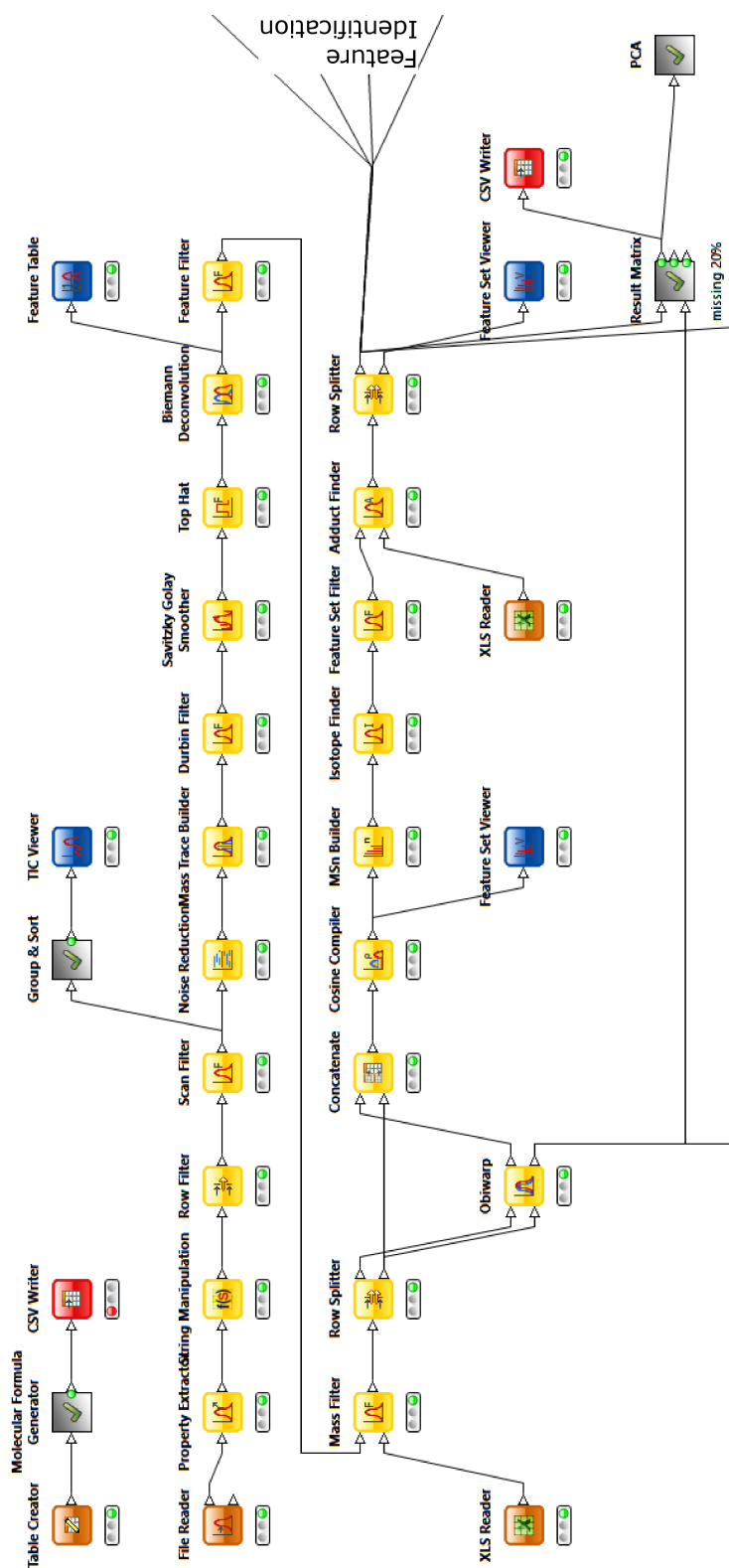


Figure 3.2: Extract of the processing workflow for metabolite identification with MassCascade in KNIME. The LC-MS/MS data is read in via the *File Reader* node before the features are extracted from the pre-filtered data. The deconvoluted features are filtered for contaminants and grouped into compound spectra through a cosine correlation matrix (*Cosine Compiler*). Isotope and adduct annotations are added before the compound spectra are queried against an in-house database ('Feature Identification', not shown).

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

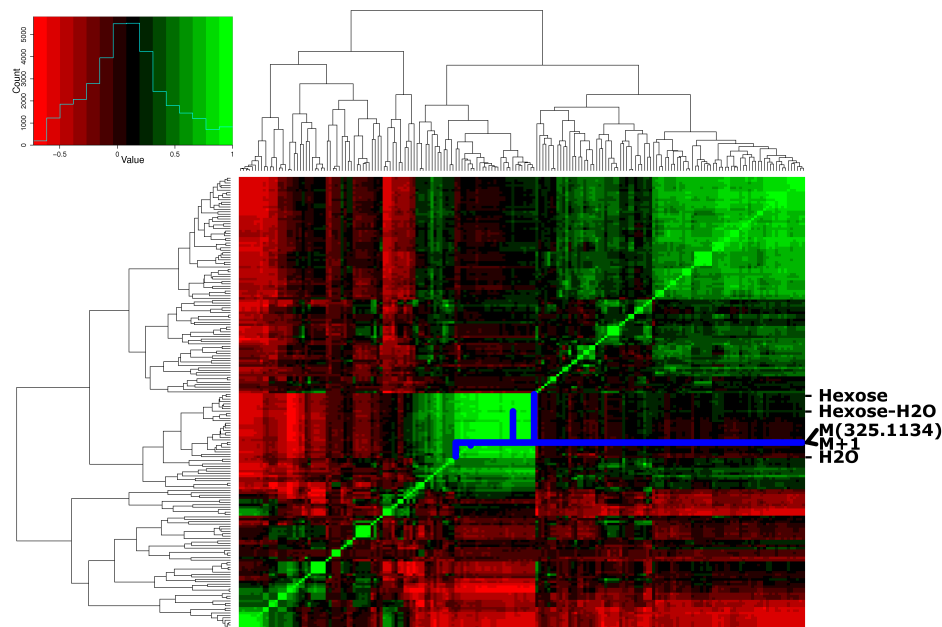


Figure 3.3: Annotated correlation heatmap of tomato study features. Pearson correlation is indicated from red to green,  $[-1, 1]$ . Rows and columns are clustered (re-sorted). They are not ordered by their respective  $m/z$  values. Annotated features such as at  $[M+H]^+ = 325.113$  (row highlighted in blue) show strong positive correlation with their putatively related features. Connections are indicated with vertical blue lines. The main feature in question elutes around 90 seconds and matches multiple sugars in web-based  $m/z$  queries, in accordance with the loss-of-Hexose and loss-of-Hexose-plus-water adduct annotations.

$DW = 2$ . The smoothed (Savitzky-Golay, third order polynomial) features were deconvoluted using a modified Biemann algorithm with a signal to noise ratio of one. Common ion traces from contaminants were removed before the features were aligned cross-sample using Obiwrap. Compound spectra were compiled from the aligned features using a cosine correlation matrix with a correlation threshold of  $\rho = 0.98$ . Master fragmentation spectra were compiled with a minimum signal occurrence of two for each parent peak. Isotope and adduct annotations were added for positive ion mode with 10 ppm mass accuracy tolerance. Figure 3.3 shows a snapshot of cross-aligned features after annotation and demonstrates the necessity for isotope and adduct annotations to subsequently de-replicate features, e.g. for statistical analysis<sup>[275]</sup>.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

The resulting features were written out in a feature matrix after gap filling had been applied. Unresolvable gaps were assigned a default intensity of 10,000 units. Missingness for the matrix compilation was set to 10%. Two different sample by feature matrices were compiled: one taking all samples and quality controls into account (QC-Matrix), and a second matrix with samples only, i.e. NOR, RIN, AC<sup>++</sup>, and CNR (Sample-Matrix).

Statistical analysis was carried out in the statistical programming environment R. Pareto scaling and total signal intensity normalization were found to be optimal after several data pre-treatment methods were tested, preserving the data structure while increasing the effect of small intensity features:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j}} \quad (3.4)$$

Here,  $x$  denotes the feature intensity in the sample by feature ( $i$  by  $j$ ) matrix and  $s$  represents the standard deviation. Missing values that could not be back-filled, i.e. values with default intensity 10,000, were imputed using a 10-component PCA model (NIPALS) before PCA was used to inspect the data and discover trends. Observing the effect on quality control samples, total signal intensity normalization was applied to remove batch variation and uni- and multivariate outliers were removed. An orthogonal partial least squares (OPLS) model with three orthogonal components was built to inspect the ripening trajectories of the tomato cultivars – characterised by their features ( $X$  data table) – over time ( $Y$  data table). The resulting model was subsequently used to identify features for identification and biological interpretation.

The selected features were run against a total of four different libraries: Syngenta’s in-house library, KEGG, ChEBI, and PubChem Compound. These libraries were used in three different scenarios in order to evaluate the scoring framework. All scenarios ran fully automated within the workflow environment. In each scenario, annotations retrieved from the individual reference libraries for all previously selected features were ranked to indicate the confidence in the retrieved annotations.

Each of the three scenarios employs a different set of filters: in the first scenario

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

no filter was applied. In the second scenario an isotope filter was added. In the third scenario a fragment filter was added. Features were queried sample by sample against the *individual* reference libraries before a ranked consensus list was built from all samples (excluding control samples) with a  $m/z$  tolerance of 10 ppm, retention time tolerance of 5 seconds, and allowed missingness of 20%. Missingness is listed as a separate filter in Table 3.1 and was applied to all scenarios when the ranked consensus list was build.

The consensus list contained the identified metabolites ranked by an averaged score based on the applied filters for isotope patterns, fragmentation ( $MS^2$ ), and adduct information that were first applied to each sample individually. The samples were first annotated and ranked individually to place higher emphasis on individual samples: external effects such as noise or batch effects could distort alignments and reduce final scores if all samples would be aligned first to generate a feature matrix. The features in the identity table are condensed to single entries per feature  $m/z$  value with averaged scores.

#### 3.4.1 Known Identification

A reference library was compiled from the KEGG pathway database and combined with an experimentally measured in-house library, which included retention times of the measured standards. The KEGG library, version 57, was downloaded in SDFfile format. Disconnected structures were reduced to their largest fragment and charged species were removed before major monoisotopic masses were calculated for query library generation. Duplicate molecule records or stereoisomers of the same compound were removed *via* SMILES line notation comparison, resulting in a reference library of 12,022 entries. Canonical SMILES for comparison were computed through the ChemAxon Marvin Extensions KNIME Feature, 2.6.3.v0135. The  $m/z$  search tolerance was set to 10 ppm. The in-house reference library contained 839 entries. The in-house library was originally measured for metabolites relevant to plant metabolomics. The KEGG pathway database covers core metabolic pathways.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

#### 3.4.2 Known Unknown Identification

The PubChem database was used as reference library and queried for the target  $m/z$  values. The PubChem database contains a range of diverse chemical compounds that are mostly irrelevant for the task of metabolite identification. To avoid the retrieval of exotic molecular structures that contain elements uncommon in metabolites, the molecular formulas calculated for the last scenario were used as filter for retrieved hits via PubChem’s web service. A total of 46,604 unique, uncharged, and connected compounds were retrieved this way and compiled as reference library.

#### 3.4.3 Unknown Identification

Molecular formulas were generated for each feature that served as input for an *in silico* structure generator (see appendix for a complete list). The molecular formulas were generated for the selected charge-corrected feature masses with a tolerance of 0.05 amu and checked for validity with the following rules:

- elements carbon, nitrogen, oxygen, and hydrogen only,
- element frequencies must not exceed 39 (C), 72 (H), 20 (N), and 20 (O)<sup>[240]</sup>,
- compliance to the nitrogen rule<sup>[276]</sup>,
- Ring Double Bond Equivalent (RDBE) values within  $[-0.5, 30]$ <sup>[248]</sup>.

Sum formulas violating one or more of the rules were removed. The structure generator Molgen<sup>[277]</sup> was run against the list of deduced sum formulas with a 30 min timeout per sum formula. To reduce computational complexity, the number of cycles was constrained to 0-2, ring-size was limited to 0-10 atoms, and the maximum number of structures was set to 50. Default values were used for all other parameters. A total of 11,159 structures were generated and compiled as reference library.



## 3.5 Results

Data processing resulted in two matrices of size 219 by 211 (QC-Matrix) and 156 by 189 (Sample-Matrix). The matrices contained 292 (0.64%) and 23 (0.07%) missing values that could not be back-filled using raw data. They were imputed through a PCA model approach.

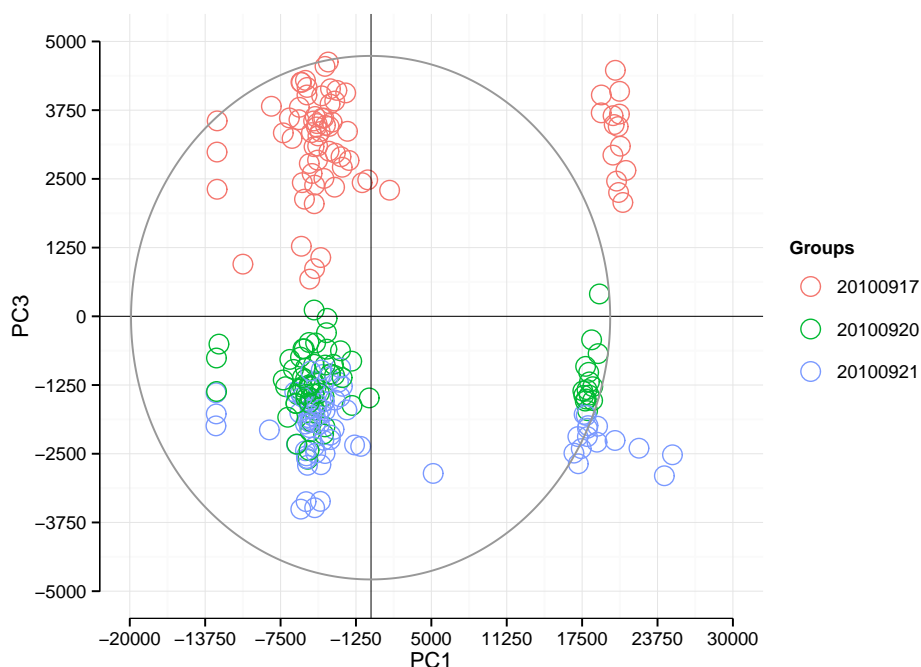
### 3.5.1 Analysis of the Quality Controls

Initial PCA on the Pareto scaled matrix shows clear batch variation in principal component 1 and 3 between samples acquired on 17.09 and 20/21.09 (Figure 3.4a). Total signal intensity normalization reduces observed batch variation but can not eliminate it completely (Figure 3.4b). The 73 affected samples acquired on 17.09 are dropped before further analysis, removing instrument induced structure. In the PCA on the trimmed data set Tomato QC samples, blanks, and pooled samples are each grouped together (Figure 3.4c). The biological samples are distributed along the third principal component. Expected clustering of Tomato QC samples, blanks, and pooled samples indicate absence of additional biological uninduced variation in the data set. Principal components 1 and 3 are chosen because they illustrate the batch effect most dramatically. Please note, that the homoscedasticity criteria was explored using 10 pooled samples. Slight heteroscedasticity is notable but could not be completely removed using power or log transformations.

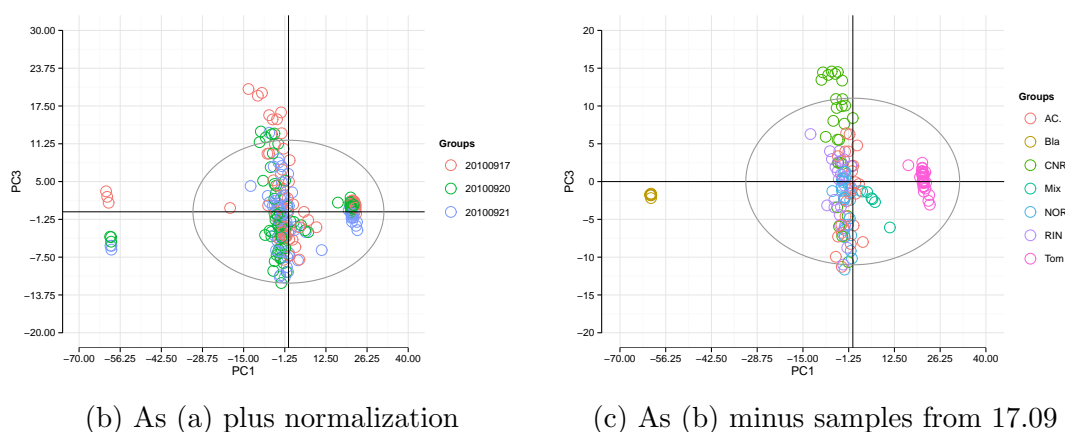
### 3.5.2 Analysis of the Tomato Samples

Following results from the QC-Matrix, samples from 17.09 are dropped from the Sample-Matrix, which contains the aligned features from the tomato samples. The matrix is total signal normalized and Pareto scaled before normality of variables is explored using a skewness measure and a visual representation in the form of quantile-quantile plots (Q-Q plots), resulting in the removal of 20 variables after inspection of the corresponding extracted mass chromatograms. Po-

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION



(a) PCA: Pareto scaled QC-Matrix (acquisition date)



(b) As (a) plus normalization

(c) As (b) minus samples from 17.09

Figure 3.4: Overview of the tomato cultivars data set based on the QC-Matrix: All samples plus blanks, Tomato QC, and pooled samples. All PCAs show the first and third principal components to illustrate batch effects. The gray circle indicates Hotelling's  $T^2$  statistic for a 95% confidence region. (a) PCA of the Pareto scaled raw data coloured by acquisition date. Blank injections, quality controls, and samples cluster but are smeared across the third component, exhibiting a batch effect. (b) As (a) with the data points normalized by total signal intensity. (c) As (b) minus samples acquired on 17 September with samples coloured by group.

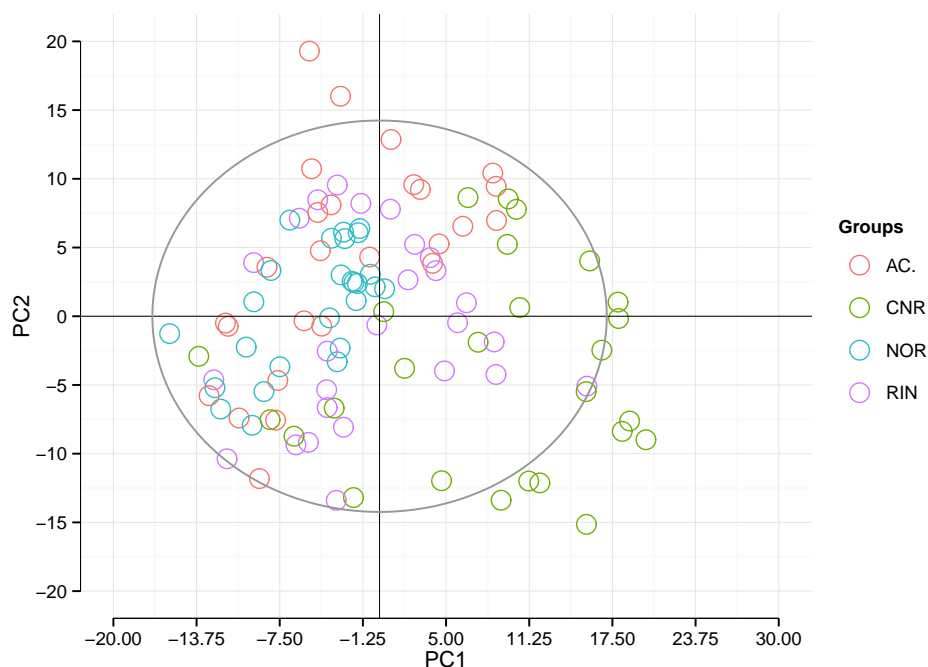
### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

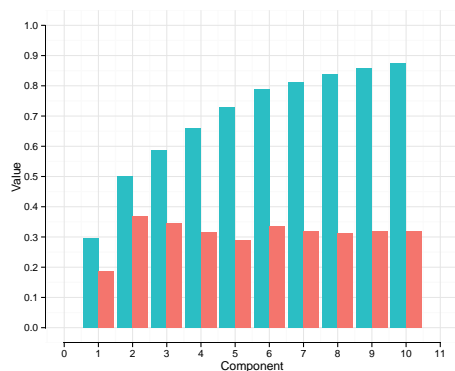
tential multivariate outliers are found using a PCA-based method by Filzmoser *et al.*<sup>[278]</sup>. After visual inspection of the chromatographic traces, all objects are kept. The resulting 10 component PCA model explains 50% variance ( $R^2$ ) in the first two principal components with a goodness of prediction of 37% ( $Q^2$ ), Figure 3.5. The coefficient of determination  $R^2$  is the total ratio of variance that is being explained by the model. The goodness of prediction  $Q^2$  results from internal seven-fold cross-validation. The applied ‘krzanowski’ type cross-validation<sup>[279]</sup> sequentially leaves out rows (features) and columns (samples) of the input matrix to build fold models that give loadings and scores respectively. Combinations of these loadings and scores are then used to estimate completely left out values (for cross-validation). The  $Q^2$  can be understood as the ratio of variance that can be predicted *independently* by the model. Samples from genotypes AC<sup>++</sup> and CNR are dominating the PCA plot. Wild type AC<sup>++</sup> samples are spread about the quadrants I, II, and III quadrant, whereas mutant CNR samples are most notably in quadrants I, III, and IV. NOR and RIN samples occupy similar space in the plot. The plot suggests that groups AC and CNR develop through separate ways. Potential outliers indicated by the Distance to Model (DModX) plot are kept because they do not exceed the critical threshold of twice the critical DModX value. Because the DModX values follow a F-distribution, the critical DModX value can be derived for a significance value of 0.05 from a F-distribution with  $j = 104$  and  $i = 169$  degrees of freedom.

To further investigate the data set and establish a valid model for feature selection and metabolite identification, an orthogonal partial least squares (OPLS) model is built (Figure 3.6). Using leave-one-out cross validation, a 1+3 model was found to be best with a bias-corrected RMSEP of 1.95. Given the nature of OPLS, the explained covariance between  $\mathbf{X}$  (the Sample Matrix) and  $\mathbf{Y}$  (sample day) is maximized for component 1. The three dimensional OPLS plot shows the ripening trajectories of all four genotypes. The trajectories of mutants NOR and RIN follow a similar path. AC<sup>++</sup> and CNR trajectories are offset and diverge after day 30 and occupy different regions along the second and third component respectively post-anthesis.

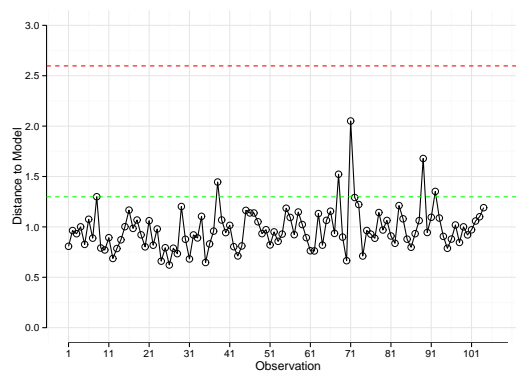
### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION



(a) PCA: Normalized, Pareto scaled Sample-Matrix minus outliers.



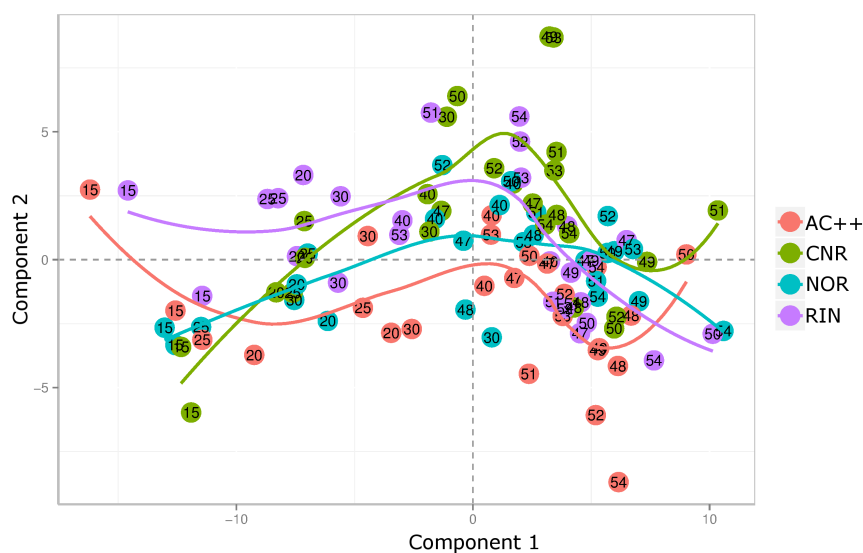
(b) Goodness of fit and prediction



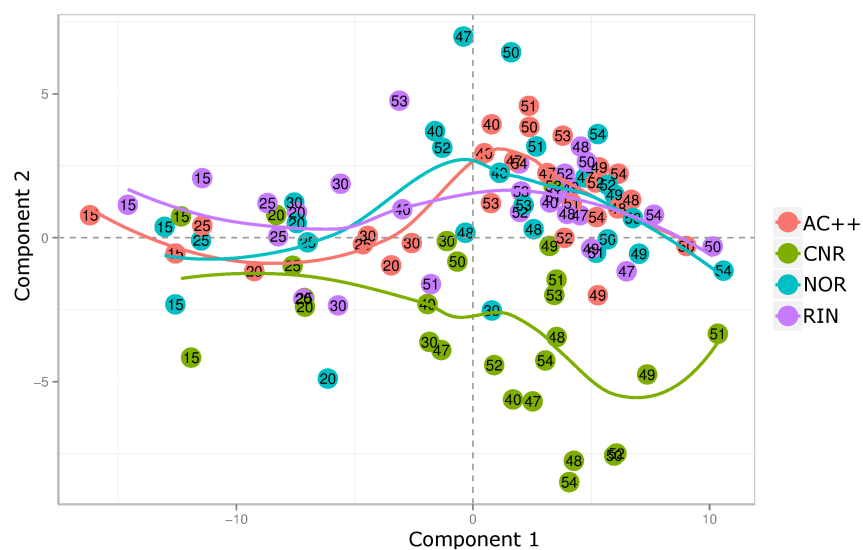
(c) Distance to Model plot

Figure 3.5: Final PCA model for the tomato cultivars data set. (a) PCA of the normalized and Pareto scaled Sample-Matrix: Genotype samples only. The samples are coloured by genotype. The gray circle indicates Hotelling's  $T^2$  statistic for a 95% confidence region. Wild type  $AC^{++}$  samples are spread about the quadrants I, II, and III quadrant, whereas mutant CNR samples are most notably in quadrants I, III, and IV. NOR and RIN samples occupy similar space in the plot. (b) Cumulative plot for the goodness of fit ( $R^2$ ) and prediction ( $Q^2$ ). (c) Distance to Model plot with a critical value threshold for potential outliers of 1.29866 (green line) and twice the critical value threshold for definitive outliers (red line).

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION



(a) OPLS: Genotype time trends (comp 1,2)



(b) OPLS: Genotype time trends (comp 1,3)

Figure 3.6: OPLS model for the tomato cultivars data set. Score plots for the first vs. second (a) and first vs. third (b) component are shown with a trendline approximated via LOESS. The samples are coloured by genotype and labelled by sample date. Covariance is maximized between the Sample-Matrix and the sample day in the first component. The trajectories of mutants NOR and RIN follow a similar path. AC<sup>++</sup> and CNR trajectories are offset and diverge after day 30 and occupy different regions along the second and third component respectively after flowering at day 47.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

Observations of mutant CNR spread significantly more than observations of wild type AC<sup>++</sup>. Subsequently separate OPLS models are built vs. time for each genotype to identify biologically relevant targets for metabolite identification. OPLS is chosen over OPLS-DA (discriminant analysis) because of familiarity with OPLS and its availability in the R environment. OPLS-DA is an extension of the OPLS technique that takes class information into account. In general, it is more suitable for discrimination modelling. Loadings of the first component of each AC<sup>++</sup> - mutant pair are plotted to single out up- and down- regulated metabolites as identified by the models. The first component encrypts **Y** – the sample time – and is thus most revealing for metabolites that change over time (Figure 3.7). Metabolites below the first quartile of the distribution of AC<sup>++</sup> and above the third quartile of the distribution of a mutant (or *vice versa*) are chosen for identification because they exhibit maximum change over time in multivariate space. That is, metabolites occupying the extreme ends of the wild type and mutant distributions are chosen so that they have high loading values in one genotype *and* low loading values in the other. This procedure results in 13, 2, and 2 singled out features for AC<sup>++</sup> - CNR, AC<sup>++</sup> - NOR and AC<sup>++</sup> - RIN models.

#### 3.5.3 Identification

A total of 13 unique features were singled out using the OPLS loadings plot approach. Missingness and element filters were included by default every time. They defined the baseline score of 160 for a missingness of 20%.

##### Known Identification

A total of 18 hits, i.e. distinct molecules, were returned. Grouped by feature, only six features had metabolite annotations. For those 18 hits, four had retention time information from the in-house library: valine, betaine, pidolic acid, and citrulline. With the exception of betaine, retention times matched within four seconds. Lists of annotations of individual features are shown in the representative Tables 3.3,

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

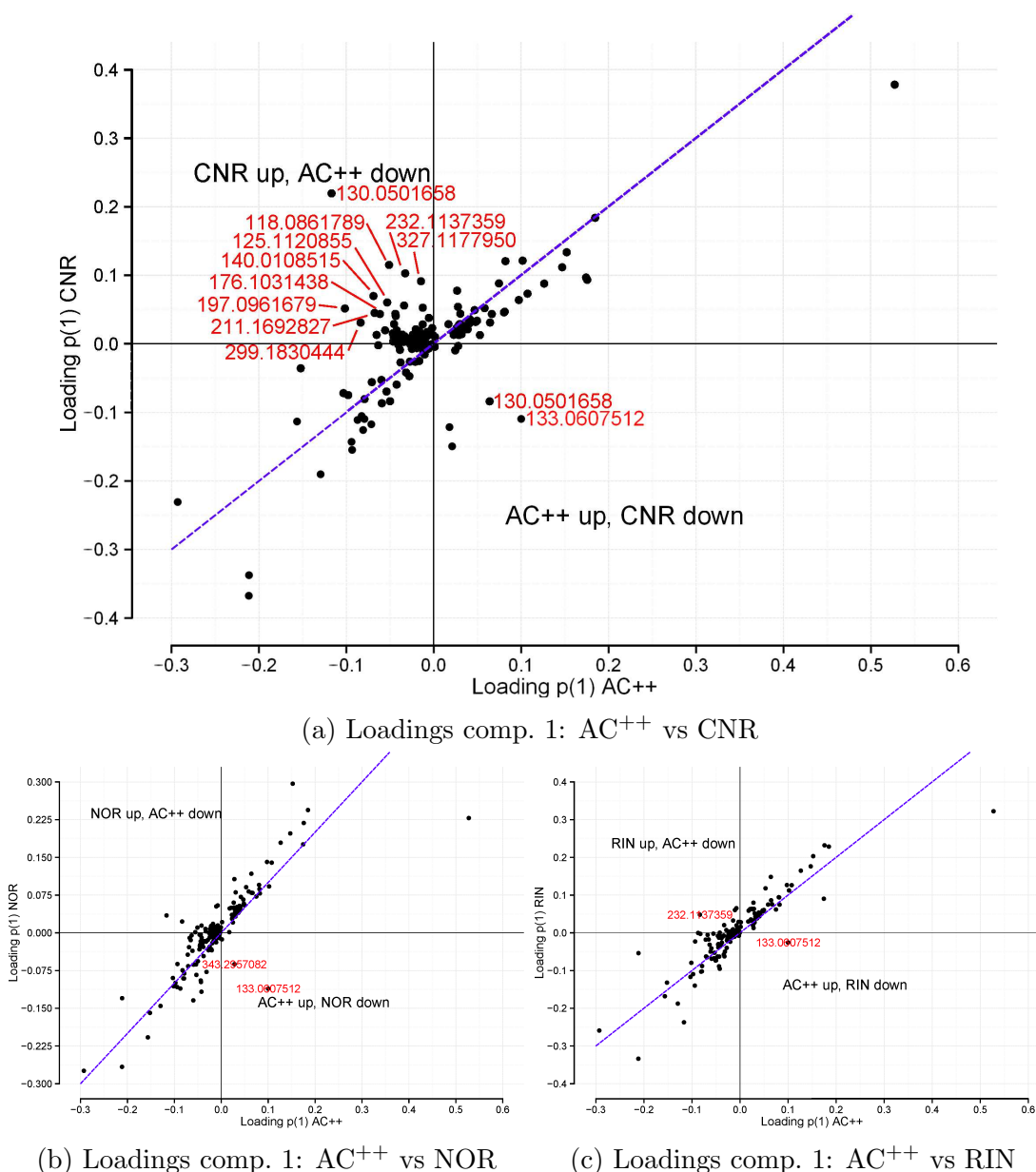


Figure 3.7: The pairwise loadings of the first component of the individual genotype OPLS models is shown for (a) AC<sup>++</sup> vs CNR, (b) AC<sup>++</sup> vs NOR, and (c) AC<sup>++</sup> vs RIN. Metabolites that do not change over time center around (0,0). The dashed blue line indicates the diagonal along which metabolite intensities increase or decrease together. Metabolites below the first quartile of the distribution of AC<sup>++</sup> and above the third quartile of the distribution of a mutant (or *vice versa*) are chosen for identification and are highlighted in red. That is, metabolites occupying the extreme ends of the wild type and mutant distributions are chosen so that they have high loading values in one genotype *and* low loading values in the other.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

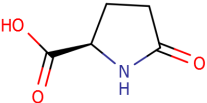
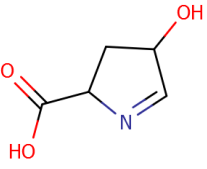
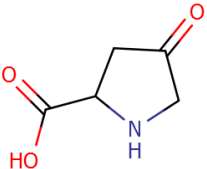
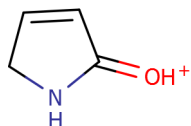
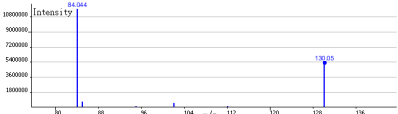
Name	Depiction	Score		
		Frag.	Isotope	None
pidolic acid <sup>†*</sup>		554	455	160
1-pyrroline-3-hydroxy-5-carboxylate		554	455	160
4-oxoproline		552	455	160
MS <sup>2</sup> spectrum	 			

Table 3.3: Metabolite annotations for  $m/z$  130.05 with cumulative scores for no, isotope, and fragment filtering. Missingness and element filters were always included. The calculated score ranks 4-oxoproline lower. All structures are highly similar. <sup>†</sup>The extracted MS<sup>2</sup> spectrum shown is dominated by a peak at 84.044 (putative structure shown). \*Matching retention time in the in-house library at 275 seconds.

3.4, and 3.5. The remaining three tables are in the appendix. The tables list the scores for the three additive filters. Where present, an extracted MS<sup>2</sup> spectrum is shown with its dominant peak annotated with a putative fragment.

The scoring framework adds discriminatory power to the ranking of metabolite annotations. The missingness and element filter set a start score of 160; the score increases if a suitable isotopic envelope is present. It increases further in the presence of MS<sup>2</sup> spectra if fragmentation signals match. The discriminatory power lies solely within the fragmentation filter. All compounds are very similar to each other. Their theoretical isotopic abundances fit the measured isotope signals and their fragmentation spectra are dominated by few signals those  $m/z$



### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

values match multiple charged fragmented substructures of each of the putative compounds. Retention time from the in-house library is the only factor that can potentially differentiate identically ranked compounds. However, only four hits from the KEGG database also had associated retention times in the in-house library. For increased confidence, all annotations would need associated retention times to enable comparisons. Betaine (Table 3.5) is the only example of a lower-ranked compound with a differing retention time.

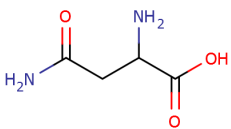
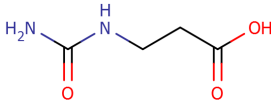
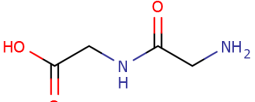
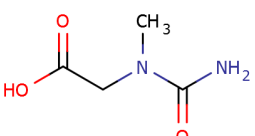
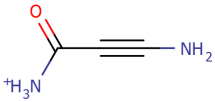

Name	Depiction	Score		
		Frag.	Isotope	None
asparagine <sup>†</sup>		351	209	160
3-ureido-propionate		350	209	160
glycylglycine		350	209	160
N-carbamoyl-sarcosine		325	209	160
MS <sup>2</sup> spectrum	 			

Table 3.4: Metabolite annotations for  $m/z$  133.061 with cumulative scores for no, isotope, and fragment filtering. Missingness and element filters were always included. The calculated score singles out asparagine. All structures are highly similar. <sup>†</sup>The extracted MS<sup>2</sup> spectrum shown at the bottom of the table is dominated by a peak at 87.055 (putative structure shown).

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

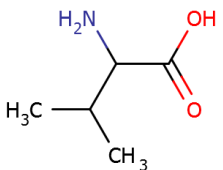
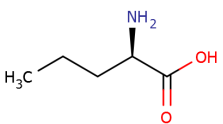
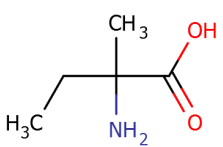
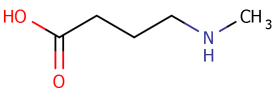
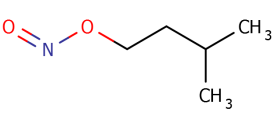
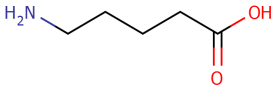
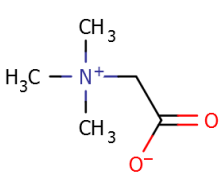
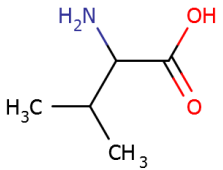
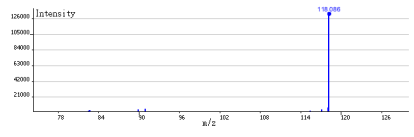
Name	Depiction	Score		
		Frag.	Isotope	None
valine <sup>†*</sup>		559	359	160
D-norvaline		559	359	160
2-amino-2-methylbutanoate		559	359	160
4-methyl-aminobutyrate		559	359	160
amyl nitrite		559	359	160
5-amino-pentanoate		557	359	160
betaine		557	359	160
MS <sup>2</sup> spectrum	 			

Table 3.5: Metabolite annotations for  $m/z$  118.086 with cumulative scores for no, isotope, and fragment filtering. Missingness and element filters were always included. The calculated score does not allow for selective ranking among the five best structures. With exception of the charged compound betaine, all structures are highly similar. <sup>†</sup>The extracted MS<sup>2</sup> spectrum shown at the bottom of the table is dominated by its parent peak (putative structure shown). \*Matching retention time information in the in-house library at 154 seconds.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

#### Known Unknown Identification

A total of 261 molecular formulas were generated. Feature 327.117 was eliminated from this scenario at the beginning because the putatively correct molecular formula  $C_{14}H_{18}N_2O_7$  was missed by the formula generator. Other formulas calculated for the same feature were removed by the isotope filter because they did not match the measured isotope pattern.

For known unknowns, i.e. hits from a generic compound database like PubChem Compound, the framework helps to narrow down the annotation space from several to a few hundred annotations but is unable to single out one compound (Table 3.6). The table lists the rank of the selected metabolite annotations from the ‘Known Identification’ scenario within the retrieved compound lists. The compound lists were grouped by score. The fraction of hits with identical scores to all hits is also listed, i.e. the number of matching references up to the group that contains the annotation divided by all matching references. For example, feature 118.086 is thought to represent valine based on the ‘Known Identification’ scenario (Table 3.5). Valine is in the highest ranked group of PubChem Compound retrievals (*Rank 1*), which includes 45% of all matching retrievals from PubChem Compound for that particular feature. The scoring framework eliminates 55% of matching compounds.

The scores are identical for all filters if the best scores from the PubChem Compound retrievals are compared to the scores from the previous scenario. Up to the isotope filter, this is to be expected because the missingness value and molecular-formula based theoretical isotopic abundances are identical. The highest ranked compounds have identical molecular formulas.

The only discriminating factor remaining is the fragmentation score. The score is also identical for all cases except for feature 197.096, which was ranked fourth: the best fragmentation score is larger by one unit. Given the information from the data, PubChem does not offer better matching compounds but a list of equally well matching compounds.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

<i>m/z</i>	Rank	Fraction	Score (PubChem/KEGG)		
			Frag.	Isotope	None
118.086	1	45%	559/559	359/359	160/160
130.05	1	68%	554/554	455/455	160/160
133.061	1	81%	350/350	209/209	160/160
176.103	1	100%	n/a	260/260	160/160
197.096	4	72%	370/369	251/251	160/160
327.117	n/a	n/a	n/a	n/a	160/160

Table 3.6: For the selected features ( $m/z$ ), the rank of the group that contains the feature’s annotation is shown and the fraction of the number of all molecules within or up to that group in relation to all retrieved hits for that feature. The top-listed annotations from the ‘Known Identification’ scenario were used for ranking. The score compares the best score of all retrieved annotations for a feature to the ‘Known Identification’ scenario. They are identical with exception of feature 197.096. Feature 327.117 did not yield the required molecular formula during the automated process.

#### Unknown Unknown Identification

The calculated average score of matching annotations is expected to be worse for the Molgen library, i.e. lower, than the average score from the ‘Known Identification’ scenario because chemical space is traversed at random.

Table 3.7 compares the best hits retrieved from the Molgen library versus previously retrieved best hits. With exception of feature 197.096, the fragmentation scores from the Molgen library are worse for all features where fragmentation spectra exist.

Feature 197.096 matches carbon-rich structures with few heteroatoms that give similar fragments if fragmented deterministically. The isotope score contributions are identical for both scenarios because identical molecular formulas resulted in the best overall score.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

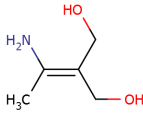
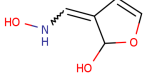
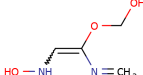
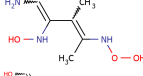
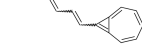
Mol. Formula	$m/z$	Depiction	Score (Molgen/KEGG)		
			Frag.	Isotope	None
$C_5H_{11}NO_2$	118.086		557/559	359/359	160/160
$C_5H_7NO_3$	130.05		553/554	455/455	160/160
$C_4H_8N_2O_3$	133.061		343/351	209/209	160/160
$C_6H_{13}N_3O_3$	176.103		n/a	260/260	160/160
$C_{14}H_{12}O$	197.096		369/369	251/251	160/160

Table 3.7: Scored metabolite annotations for the best hits in the Molgen library versus the best hits in the KEGG library for five selected ion traces. Cumulative scores for no, isotope, and fragment filtering are shown with Molgen/KEGG scores. Missingness and element filters give a base score of 160. With the exception of feature 197.096, fragmentation scores from the ‘Known Identification’ scenario are greater than scores from the Molgen library.

#### Feature Analysis

Putative feature identifications were further investigated through biological interpretation based on univariate statistics. Feature abundances across samples, their distributions, and receiver-operator characteristic (ROC) curves were compiled for every genotype.

The feature abundances shown in the scatterplots are sorted by genotype and harvest date in ascending order. The trend for the features’ abundances is indicated through locally weighted scatterplot smoothing. Boxplots show the median (middle line) and lower (25%) and upper (75%) percentiles of the abundances.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

The ROC curves describe the power of the respective features for binary classification of either of the three mutant genotypes (RIN, NOR, CNR) versus the wild type AC<sup>++</sup>. This statistic supplements the boxplots, indicating how characteristic the total change of a feature is for a genotype compared to AC<sup>++</sup>.

A complete list of univariate feature statistics for all 13 unique features extracted from the OPLS loading plots can be found in the appendix, including Pearson's correlation matrices for each genotype. Here, the six features that have annotations are discussed because of biological interpretability.

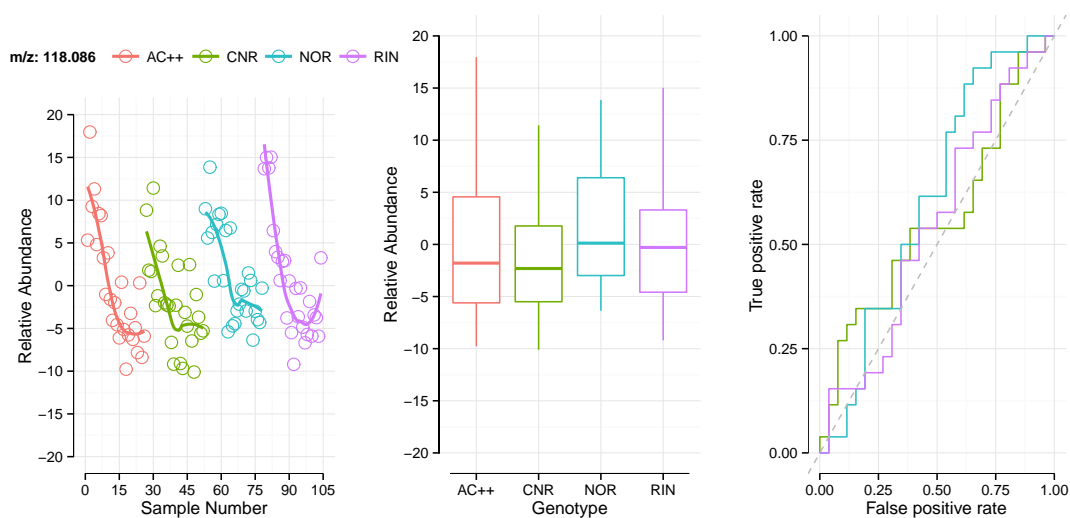
Figure 3.8a summarizes feature 118.086. The relative abundances decline during the ripening period in all genotypes. The distributions are similar and the feature does not appear to have discriminatory power in univariate space. The importance assigned by the OPLS loadings plot between AC<sup>++</sup> and CNR is not reflected.

Feature 130.05 shows a distinct difference between the AC<sup>++</sup>/CNR and NOR/RIN genotypes (Figure 3.8b). The feature abundances are not decreasing over time to the extent of the wild type and colorless non-ripe feature abundances as indicated by the trend lines.

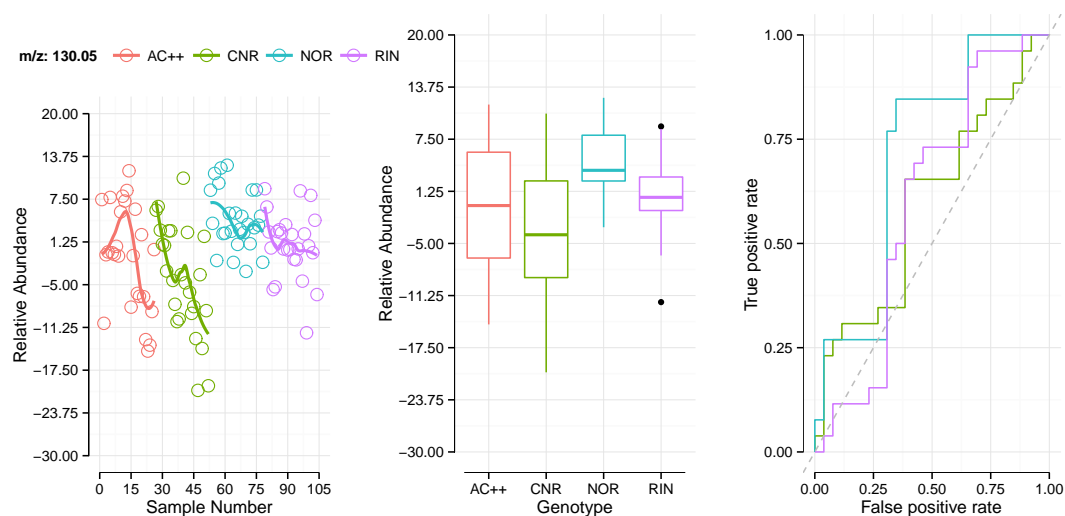
In Figure 3.9a, AC<sup>++</sup>'s level of feature 133.061 rises quicker over time than for the other genotypes. The overall difference is minor and does not impact significantly on the ROC curve. The differing behaviour is captured in the OPLS loading plots: in all three comparisons, the loadings do not indicate correlation.

Feature 176.103 (Figure 3.9b) dramatically drops in abundance for CNR, distinguishing CNR from the wild type. In Figure 3.10a and 3.10b, the levels of features 197.096 and 327.118 are spiking compared to the other genotypes.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION



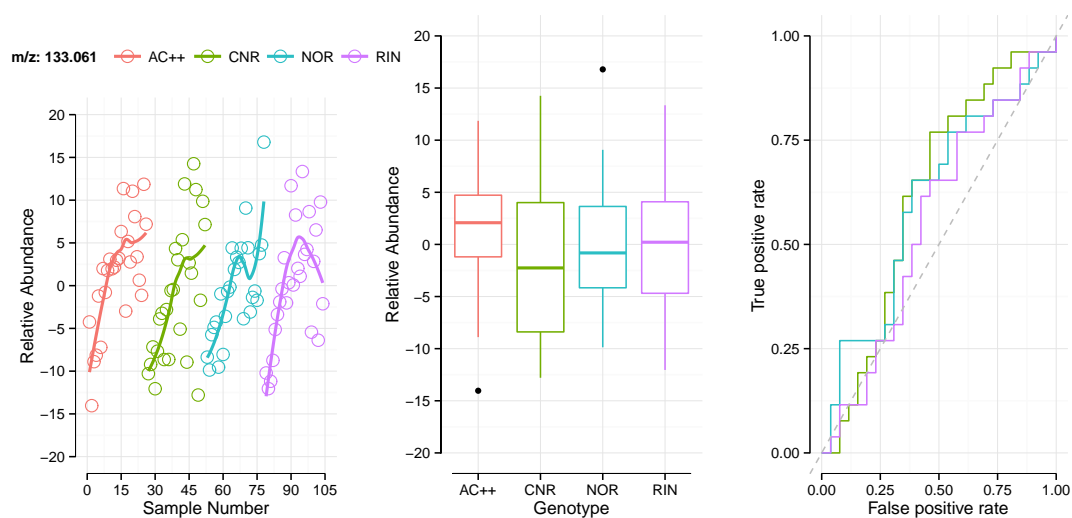
(a) Univariate statistics for feature 118.086



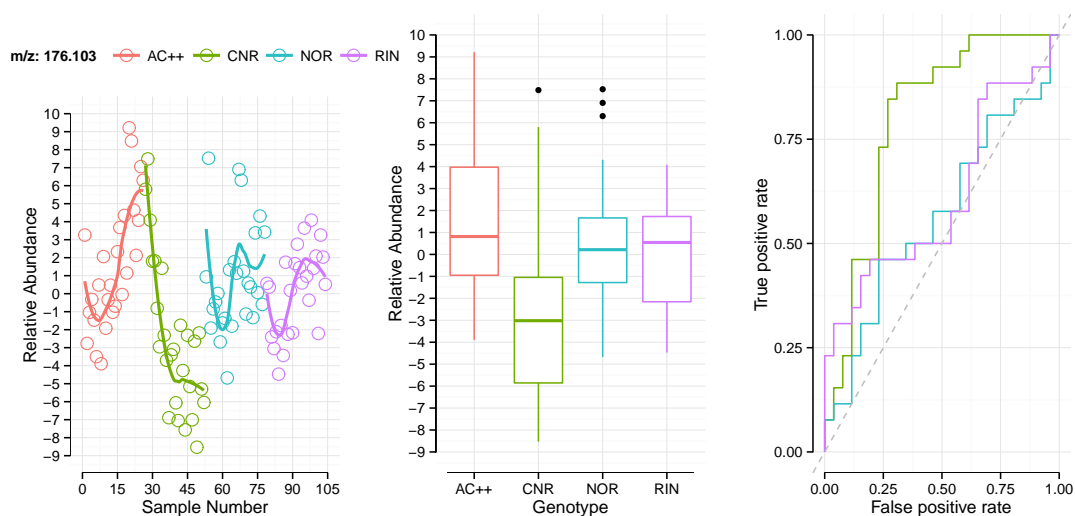
(b) Univariate statistics for feature 130.05

Figure 3.8: Univariate statistics for features 118.086 (a) and 130.05 (b). The scatterplot shows feature abundances grouped by genotype and sorted by harvest date in ascending order with applied locally weighted smoothing. The boxplots display the median (middle line) and lower (25%) and upper (75%) percentiles of the abundances. The ROC curves indicate the binary classification power of the mutant genotypes versus the wild type AC<sup>++</sup>.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION



(a) Univariate statistics for feature 133.061

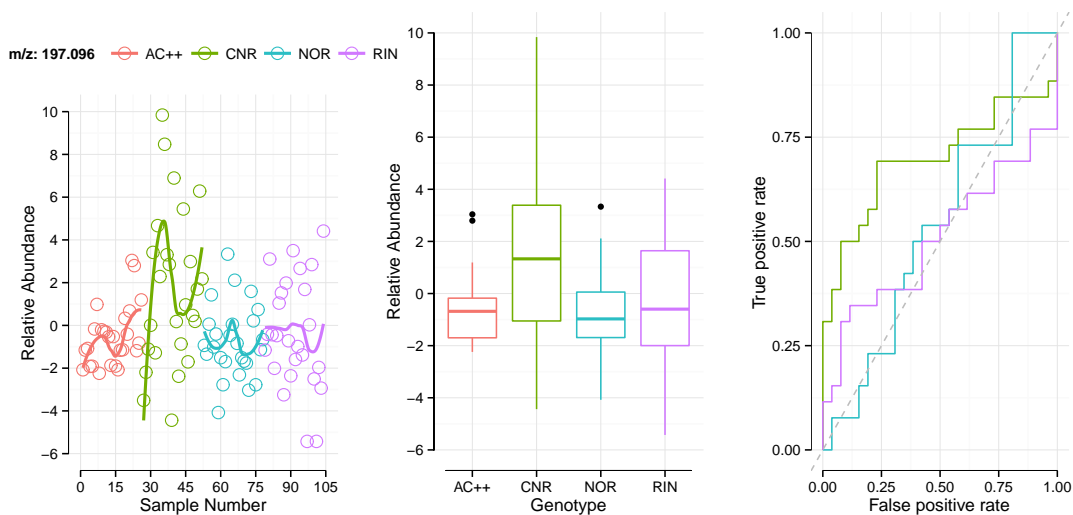


(b) Univariate statistics for feature 176.103

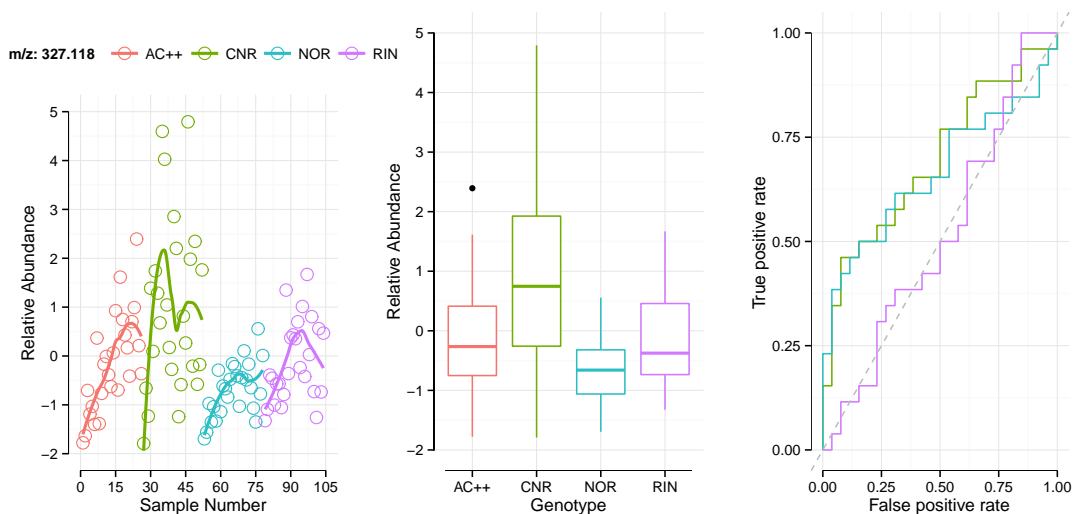
Figure 3.9: Univariate statistics for features 133.061 (a) and 176.103 (b). The scatterplot shows feature abundances grouped by genotype and sorted by harvest date in ascending order with applied locally weighted smoothing. The boxplots display the median (middle line) and lower (25%) and upper (75%) percentiles of the abundances. The ROC curves indicate the binary classification power of the mutant genotypes versus the wild type AC<sup>++</sup>.



### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION



(a) Univariate statistics for feature 197.096



(b) Univariate statistics for feature 327.118

Figure 3.10: Univariate statistics for features 197.096 (a) and 327.118 (b). The scatterplot shows feature abundances grouped by genotype and sorted by harvest date in ascending order with applied locally weighted smoothing. The boxplots display the median (middle line) and lower (25%) and upper (75%) percentiles of the abundances. The ROC curves indicate the binary classification power of the mutant genotypes versus the wild type  $AC^{++}$ .

## 3.6 Discussion

Thirteen features were selected for identification that were believed to be contributing to differences in ripening between the investigated genotypes. Three scenarios were applied to test the scoring framework. Score contributions from the missingness, adduct, and isotope filters added little discriminatory power to the ranking. These filters primarily ensured that highly unlikely compound annotations were removed from the beginning.

The remaining list of similar compounds, i.e. identical molecular formulas and presence/absence of cycles, still comprised hundreds of entries for the ‘Known Unknown’ scenario (using the PubChem Compound database) that could not be trimmed down further using fragmentation information. This can partially be explained by poor fragmentation behaviour of the measured ions, e.g. due to insufficient fragmentation optimization in the instrument, and deficiencies that result from a deterministic structure fragmentation approach that traverses all possible fragments without physical constraints to rank and discard unlikely fragments.

For the smaller, biologically relevant database used in the ‘Known Identification’ scenario, the score contribution from the fragmentation filter allowed the discrimination of similar retrieved hits highlighting a single or a few compounds. The score differences are small but reflect the absence of explainable fragmentation signals that are important for identification. Currently, the structure fragmentation tool, which is used in the fragmentation filter, does not predict abundances of fragments. It only generates fragments and abundances are set to one. Because the fragmentation filter calculates a score based on the match of the predicted to the measured fragmentation spectrum that takes abundances into account, a more elaborate fragmentation tool that also predicts abundances is likely to increase the scoring differences seen. This is exemplified in the ‘Unknown Unknown Identification’ scenario. If chemical compounds for annotations were computed randomly, the scores of the best matching annotations were smaller than for the relevant annotations from the ‘Known Identification’ scenario.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

To consolidate the ranked feature annotations and demonstrate correct data processing and identification by MassCascade, univariate analysis was carried out. The features were initially selected based on a multivariate OPLS approach. Because univariate statistics do not capture co-linearity between variables, not all features appear to be of importance in their univariate evaluation. The wild type AC<sup>++</sup> and the mutant genotypes CNR, NOR, and RIN have been shown to share the same compounds but that their concentrations vary during ripening, resulting in the different trajectories captured in Figure 3.6<sup>[280]</sup>. RIN and NOR mutants never achieve a ripe stage, i.e. their ripening stops before maturity. This behaviour is also captured in their metabolic profile as can be seen by the OPLS plot and in the abundances of feature 130.05 (Figure 3.8b), where pidolic acid has been reported to be closely correlated to ripening<sup>[281]</sup>. Asparagine concentrations, feature 133.061 (Figure 3.9a), have been reported to increase during ripening<sup>[280]</sup> (major N form in plants), whereas valine concentrations, feature 118.086 (Figure 3.8a), have been reported to decrease during ripening<sup>[282,283]</sup>. Citrulline is closely linked to its precursor arginine, which is regarded as plant growth regulator in the greater network of polyamine-mediated effects<sup>[284]</sup>, potentially explaining its complete defect in the CNR genotype. The discussed features identifications and extracted abundances appear to be reasonable within the outlined biological context.

## 3.7 Conclusion

Knowledge-based compound identification can help to reduce the size of lists of putative metabolite annotations in metabolomics studies. The implemented approach of information aggregation followed by scoring and ranking based on adduct, isotope, and fragmentation data helps to discard biologically irrelevant structures as can be seen in the example scenarios. Annotations could be removed after initial  $m/z$ -based assignment.

Similar structures retrieved from compound databases are more difficult to differentiate based on mass spectrometry data alone as demonstrated in the next

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

section – unless multiple MS<sup>n</sup> are recorded, a currently infeasible approach for untargeted studies. As the ‘Known Identification’ scenario indicates, poor fragmentation behaviour – due to low ionisation efficiency – and low  $m/z$  values render discrimination, e.g. between valine/norvaline or pidolic acid/1-pyrroline-3-hydroxy-5-carboxylate impossible. Confidence in metabolite rankings can be increased by addition of retention time (or retention index) information. However, this approach is limited by the poor ratio of the maximum feasible size of any measured in-house library versus the tangible metabolite space. In general, the addition of orthogonal sources of information increases the ability to rank putative metabolite annotations.

The study design, data, and processing/analysis processes are captured using open tools to achieve maximum transparency and reproducibility in line with current community-wide efforts in metabolomics. To that end, the complete study including the in-house library has been deposited in the MetaboLights database and a Scientific Data article was submitted to the Nature publishing group. The strength of MassCascade, introduced in chapter 2.1, has been demonstrated by using it for data processing and analysis including metabolite annotation and ranking. The created workflow that was used throughout the analysis process is available online, can easily be shared, and includes all set parameters. Its integration into the workflow environment KNIME enabled additional functionality, such as the dereplication of features or the required data transformation for the compilation of reference libraries. Open data or the formation of a ‘Datument’ has been thoroughly achieved.

Outstanding problems involve the generation of rational fragmentation spectra *in silico* that also give predicted signal abundances. This will significantly increase the power of the scoring framework for annotations from highly similar molecular species.

## 3.8 Technical Validation

The identification framework implemented in MassCascade – previously applied to the analysis of the tomato samples – is validated against a publicly available benchmark data set to further consolidate the methodology. The *Critical Assessment of Small Molecule Identification* (CASMI) is an inaugural open contest based on a common open dataset to evaluate identification methods for LC- or GC-MS data. CASMI was founded by Emma L. Schymanski and Steffen Neumann in 2012 and is at the time of writing in its third round (2014)<sup>[285]</sup>.

Test data has been taken from the CASMI contest in 2013 for which the challenge data and the solutions are available online (<http://www.casmi-contest.org/2013/>). For a comprehensive description of the data format, ranking procedures, and structure of the contest, please see the article by Schymanski *et al.*<sup>[286]</sup>.

A total of 11 out of 16 challenges are selected from category 2 (best structure identification). The missing five challenges are centred around data acquired in negative ion mode, currently not fully supported by the identification framework in automated mode. ‘Automatic method’ is a boolean flag in the CASMI contest indicating whether the metabolite identification process has run without manual intervention, the scenario chosen for this technical validation. Category 1 (best molecular formula) has been skipped because it is an inherent part of category 2. Following CASMI reporting standards, details per challenge are listed in Table 3.8. For details on the individual rankings, please see Schymanski *et al.*<sup>[286]</sup>.

### 3.8.1 Methods

Test data is provided as peak lists in flat files. Each challenge contains one peak list for MS and a separate peak list for MS/MS. No raw data or chromatographic profiles are provided. The accompanying meta files contain the instrumental parameters plus a brief description of the challenge including tips such as ‘contains aromatic structures’ or ‘contains amide bonds’.

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

For each challenge, both MS and MS/MS peak lists are converted to MassCascade Features and aggregated in a Feature Set. Fragmentation spectra are linked to their parent peaks either through information provided in the respective meta files or through manual evaluation of the MS and MS/MS spectra. The generated Feature Sets serve as starting point for the annotation and identification process.

Adduct and isotope annotations are calculated using a 10 ppm  $m/z$  tolerance window before the PubChem Compound database is queried through the ChemSpider web service for putative structures for the parent peak in MS<sup>1</sup>. Simultaneously, the MS and MS/MS spectra are queried against MassBank with default parameters. Mass accuracies were adjusted between 3 and 10 ppm guided by the challenge meta files. Retrieved metabolite annotations are submitted to brute-force feature fragmentation and MS/MS assignment before candidate ranking is carried out using all score filters (default parameters).

MassBank spectra deposited by 'CASMI2013 organizers' are ignored in this exercise. Otherwise, in the case of matching MassBank MS/MS spectra, the fragmentation filter's score is replaced by half of the score of MassBank's query score for the fragmentation spectrum, ensuring that both scores are comparable. Ranked candidates are further filtered using a SMARTS substructure search based on the challenge tips that describe structural properties of the correct solution.

#### 3.8.2 Results & Discussion

Parsing and formatting the plain data files were the most time-consuming steps. The annotation and identification workflow ran fully automated after initial setup and deduction of the SMARTS query strings for each challenge. The ChemSpider and MassBank web services were reliable but slow in execution. Preference was given to queries against MassBank over Pubchem Compound because MassBank is an experimental mass spectral database and thus able to provide better confidence in putative metabolite annotations.

A total of one MS spectrum with a single associated MS/MS spectrum was gen-

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

#	rank	tc	bc	wc	ec	rrp	p	wbc	wwc	wec	wrrp
1	7	2954	0	2947	7	1.00	0.00	0.00	0.99	0.01	0.99
2	1	1278	0	1277	1	1.00	0.00	0.00	1.00	0.00	1.00
4	226	323	0	90	226	0.65	0.00	0.00	0.08	0.92	0.08
5	1	237	0	236	1	1.00	0.00	0.00	1.00	0.00	1.00
7	22	44	0	22	22	0.76	0.04	0.00	0.26	0.86	0.14
8	3	17	0	14	3	0.94	0.14	0.00	0.58	0.42	0.58
9	1	63	0	62	1	1.00	0.02	0.00	0.98	0.02	0.98
10	6	18	1	12	5	0.82	0.12	0.12	0.31	0.58	0.31
13	366	384	0	18	366	0.52	0.00	0.00	0.01	0.99	0.01
14	997	3457	0	2460	997	0.86	0.00	0.00	0.36	0.49	0.51
16	–	1879	–	–	–	–	–	–	–	–	–

Table 3.8: Results of the technical validation using the CASMI data set of 2013. Results are reported following the CASMI reporting standards: *#*, challenge; *rank*, absolute rank of correct solution; *tc*, total number of candidates; *bc*, number of candidates with a score better than correct solution; *wc*, number of candidates with a score worse than correct solution; *ec*, number of candidates with same score as the correct solution; *rrp*, relative ranking position (1.0 is good, 0.0 is not); *p*, score of correct solution; *wbc*, sum of scores better than correct solution; *wwc*, sum of scores worse than correct solution; *wec*, sum of scores equal to correct solution; *wrrp*, RRP weighted by the scores (1 is good).

erated for each challenge. The result of the analyses is summarised in Table 3.8. Note that challenges 3, 6, 11, 12, and 15 are missing because of issues around the negative ion mode as explained in the preceding section. Because the filters iteratively remove irrelevant candidates, e.g. disconnected structures or structures that do not match measured isotopic envelopes, and do not report complete lists of all candidates, removed candidate entries were given a base score of 200 in order to calculate the normalised scores (*p*, *wbc*, *wwc*, *wec*, and *wrrp*) through the total sum of scores as described by Schymanski *et al.*<sup>[286]</sup>. The base score reflects the score that any candidate would be given by default if it passed the missingness filter. This can be assumed in this exercise because the data reflects extracted features of interest.

Similar to other contestants of the CASMI challenge, we used MassBank and Pub-

### 3. KNOWLEDGE-BASED COMPOUND IDENTIFICATION

---

Chem Compound as query databases to retrieve collections of metabolite structures as starting point. The results table shows that the identification framework was able to extract the correct structures from larger sets (up to 3,500) and rank them within the top 10 candidate structures in 6 out of 11 cases. The correct candidate structure was ranked lower in challenges #7 (rank 22), #4 (rank 226), #13 (rank 366), and #14 (rank 997). The correct solution was not found at all in challenge #16.

The solution to challenge #7 is a pentameric proanthocyanidin (Cinnamtannin A3). Similar structures such as other flavanol-based compounds could not be differentiated from the correct solution due to the repetitive nature of polymeric compounds, resulting in 22 equal ranked candidates. In challenges #4, #13 and #14, the high rank also results from limitations of the framework to differentiate very similar molecular structures, resulting in 226, 366 and 997 equal ranked candidates. In challenge #16 the correct structure was removed by the isotope filter: the measured isotopic peak ( $[M + 1]^+$ ) did not fit the value calculated by the program.

Overall, an evaluation of the absence of better candidates in all but one challenge combined with the large numbers of equal ranked candidate structures, indicates the difficulty in metabolite identification using mass spectrometry data alone. Whereas – with regard to the CASMI contest – the framework’s performance resides between the (semi-)automated methods of the contestants ES and FA (results of the contest are available on the official website), the framework performed poorly compared to the contestants who used more manual methods of investigation and took additional information into account such as species, retention times, and classifications such as ‘natural product’.



# COMPUTATIONAL WORKFLOWS FOR CHEMINFORMATICS

---

## 4.1 Introduction

The routine work of a cheminformatician involves the processing of collections of small molecules. Standardising molecules, e.g. adding hydrogens or removing unconnected structures, calculation of molecular descriptors, and visualisation of chemical structures in two- or three-dimensional space are just a few examples of recurring tasks that are carried out upstream of cheminformatic pipelines. Several free and open source cheminformatics libraries and tools have been developed to deal with these tasks, such as the CDK<sup>[287]</sup>, RDKit<sup>[191]</sup>, and OpenBabel<sup>[193]</sup>.

Building a comprehensive pipeline for handling small molecules requires a basic understanding of a scripting language to concatenate input and output from different tools or call functions from a cheminformatics library. For experimental scientists, usage of APIs (application programming interfaces) or programming languages can add a constraint to more in-depth analysis. Standalone tools suffer from limited scope, i.e. they mostly do one thing only. Even simple tasks like the visual characterisation of a chemical library<sup>[288]</sup> require importing and exporting

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

of data in various formats using different tools. This poses challenges related to the maintainability and modularity of custom pipelines that are created for these tasks.

Workflow environments circumvent the above mentioned challenges to various degrees by providing a common platform for different tools and have become popular with the science community<sup>[289]</sup>. A cheminformatics plug-in, KNIME-CDK, has been developed based on the Chemistry Development Kit (CDK), an open-source cheminformatics library. It wraps elements of the library’s core functionality and exposes it to the user. In contrast to other cheminformatics plug-ins available in KNIME, the project and its core library are fully open and community-driven.

KNIME-CDK is part of KNIME’s community contributions and adheres to their software versioning system. Separate versions exist for the major KNIME releases, 2.7, 2.8, and 2.9, and for the nightly version for active development. KNIME-CDK is an official ‘Trusted Community Contributions’ since version 2.9.

### 4.2 KNIME-CDK’s Implementation

KNIME-CDK has been developed in Java<sup>®</sup> 1.6 for the KNIME legacy version 2.6 and Java<sup>®</sup> 1.7 for all KNIME versions greater than 2.6. Following KNIME’s data model, individual CDK molecule representations are stored in their own data cell type, the atomic unit for tabular data transfer from one node to another. Community cheminformatics plug-ins come with their own cell types that uniquely capture the underlying library’s molecule representations. To make data cell types of other cheminformatics plug-ins accessible without explicit conversion on the code level, generic KNIME chemistry types are used as buffer in between cheminformatics plug-ins. Thus, KNIME-CDK depends on the generic KNIME chemistry types that serve as a wrapper for common chemistry file formats (see section 1.2.2).

## 4. WORKFLOWS FOR CHEMINFORMATICS

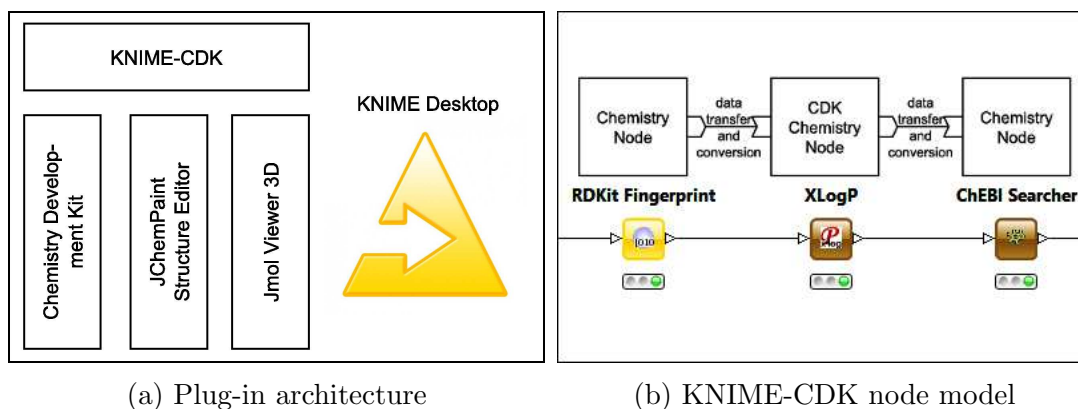


Figure 4.1: Overview of the KNIME-CDK architecture and node model. (a) The plug-in forms part of the KNIME Desktop and serves as an interface between the CDK core library for data processing, JChemPaint for chemical drawing, and Jmol for 3D visualisation. (b) The node model enables molecular compounds from any generic chemistry format to be processed within the KNIME-CDK environment. A compound from the RDKit cheminformatics package is processed in CDK (logP calculation) before it is further used for a web query that does not form part of the CDK plug-in.

### 4.2.1 Structure

The KNIME-CDK plug-in follows the classical KNIME node architecture as described in section 1.2.4. Nodes run within Java's<sup>®</sup> concurrency framework<sup>[229]</sup> with threading enabled for fast execution. KNIME-CDK uses CDK for data processing and visualisation, JChemPaint<sup>[290]</sup> for chemical drawing, and Jmol<sup>[291]</sup> for visualisation in 3D space (Figure 4.1a). No direct file input or output nodes exist. Instead, KNIME tables become usable within the KNIME-CDK environment after chemistry cell type conversion from a generic format to the internal KNIME-CDK representation (Figure 4.1b).

### 4.2.2 Persistence

Data persistence is guaranteed via the Chemical Markup Language (CML)<sup>[207]</sup> serializing the molecule when necessary. The underlying CDK molecules are handled and stored within data cells in standardized form, i.e. with implicit hydrogen atoms added, atom types perceived, and aromaticity detected. This guarantees

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

consistency across all nodes and simplifies usability of the plug-in by automating technical details from the user, hence allowing the scientist to focus on the task at hand.

The nightly build features persistence via the line notation SMILES based on a recent release of CDK version 1.5.4. In contrast to CML, SMILES – as line notation system – does not allow for additional information to be stored directly. Consequently, atom types and aromaticity are perceived every time a molecule is requested for processing, i.e. at every node. This overhead is balanced out by advantages of the SMILES system: information density and applicability within the workflow environment. Two- and three-dimensional coordinates as well as additional information about substructure highlights (see section 4.3.3) are stored in vectorised form with the SMILES notations. It should be noted that a non-canonical version of SMILES is used to increase execution speed: isomeric SMILES. However, conversion rounds of SMILES to native CDK representation to SMILES always result in the same SMILES and native CDK representation because the atom order is preserved in a separate auxiliary array. The array maps the order of the atoms in the CDK representation to the order in which the SMILES string is parsed. SMILES and auxiliary information are serialized to disk in a byte stream.

### 4.3 KNIME-CDK's Functionality

The plug-in includes methods for the generation of two- and three-dimensional coordinates, atom signatures, common fingerprints, e.g. MACCS and Pubchem Compound, two- and three-dimensional molecular descriptor values including XLogP, Lipinski's Rule of Five, offers chemical name to structure conversion via the webservice OPSIN<sup>[196]</sup> and SMARTS or substructure search abilities. The substructure search features exact stereo and charge matching unique to the KNIME-CDK implementation. In Figure 4.2 a chemical library is filtered for molecules containing a phenol group before successive hydrogen acceptor and donor count while being used for MACCS fingerprint and atom signature gene-

## 4. WORKFLOWS FOR CHEMINFORMATICS

ration. The out-port view, i.e. the resulting data table, is shown for the *Atom Signatures* node.

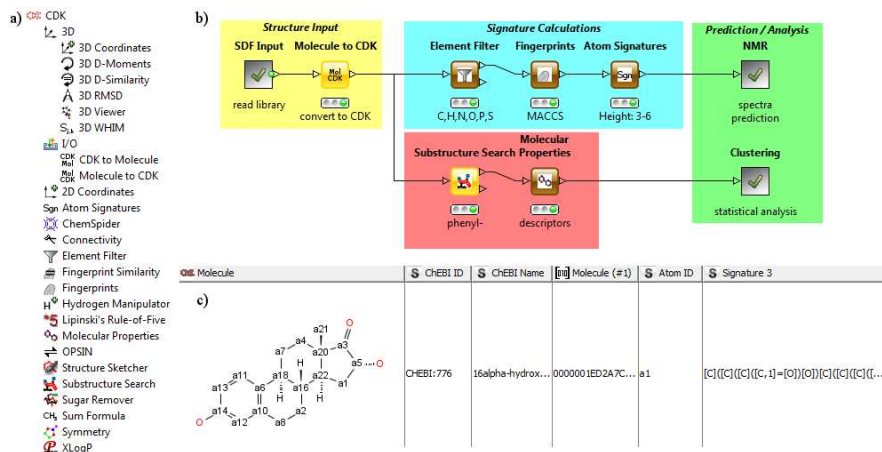


Figure 4.2: Screenshot of a KNIME-CDK workflow. a) View of the node repository showing all available nodes. b) Example workflow for descriptor calculation. The molecule library is read in and filtered for structures containing phenol groups before counting the number of hydrogen donors and acceptors (lower path). Simultaneously, MACCS fingerprints and atom signatures are calculated for the atom-filtered molecules (upper path). c) Example row from the out-port view of the *Atom Signatures* node showing the CDK molecule followed by the ChEBI identifier, name, MACCS fingerprint, atom identifier and corresponding HOSE code.

A part of AMBIT's<sup>[292]</sup> functionality – a software for cheminformatic data management based on CDK – has additionally been added to KNIME-CDK to further extend its uses. AMBIT's tautomer generator enumerates tautomeric constitutions in a rule-based fashion<sup>[293]</sup> and has also been made available as node in the plug-in: either all viable tautomers or the single best tautomer (measured in electron volt indicating the energy score) can be generated. This node helps significantly with molecule library standardization and demonstrates the use of workflow environments where multiple tools can be pooled together.

### 4.3.1 Input/Output

The plug-in accepts molecules in CML, SDFfile, MDL Mol, InChI, and SMILES formats<sup>[200]</sup> via the *Molecule to CDK* node or directly without explicit conversion.

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

The plug-in cell type's back-end supports adapter values. These adapter values ensure automatic chemistry cell type conversion between formats. *Vice versa*, CDK cell types can be written out using the same mechanisms. From KNIME version 2.9, the original `CDKCell` has been replaced with `CDKCell2` in the nightly build and cell creators and converters have been updated accordingly. The original data cell implementation is based on a `BlobDataCell` for large binary objects, significantly slowing down execution because of disk I/O overhead: blob data cells are registered and buffered separately within the KNIME framework. The newer `CDKCell2` implementation extends a regular `DataCell` instead, avoiding unnecessary separate cell handling for small objects like SMILES strings.

Alternative ways of molecule input include the structure editor `JChemPaint` (*Structure Sketcher*) and implemented web services that return molecular compounds. The `JChemPaint` project is not actively maintained at the moment but the project has been updated to the CDK version used in KNIME-CDK. It is now maintained alongside the plug-in.

On data input, molecular compounds are standardised. This includes, unless already present, generation of two dimensional coordinates, perception of atom types, addition of implicit hydrogens, and detection of aromaticity. This ensures consistency within the CDK environment independent of the processing functions applied on a compound. For example, not every node requires aromaticity to be perceived, opening up the possibility to perceive aromaticity on demand only. This introduces the possibility of subtle errors that outweigh any gain from omitting resource-consuming steps at chemistry cell type conversion. Additionally, a two-tier hash code is used for fast molecule comparisons. Given the overhead of hash code generation, a simple hash code based on just the molecular graph skeleton of depth eight and assigned charges is generated on read-in for initial comparison. If two hash codes match in a molecule comparison, a more detailed 'comprehensive' hash code is calculated for in-depth comparison based additionally on stereo-, isotope-, and radical-information.

CDK cells can be copied as cleaned CML. CDK-specific CML elements from the internal representation are removed, resulting in XML compliant to CML specifications. For the nightly build, copy actions result in SMILES copies.

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

### 4.3.2 Processing

Compounds can either be used for property calculations, which are appended to the cell row, or manipulated directly. If a compound is manipulated, e.g. by stripping salts, the original compound representation is replaced. For combination with adapter values for automatic chemistry cell type conversion, if nodes from different packages are used in direct succession, the following points need to be noted:

- The original chemistry type does not change unless the chemical compound changes.
- Property calculations are always based on the chemistry type of the plug-in which function is applied. This can go unnoticed on the node I/O level.
- If the chemical compound changes, the chemistry type of the current package is used in the output.

### 4.3.3 Visualisation

KNIME-CDK adds the CDK renderer to the set of KNIME chemistry renderers. The default renderer for chemical compounds in node tables can be chosen in the general preferences. The renderer has been adapted from the basic CDK renderer to display annotations including radicals, isotope numbers, and atom identifiers (Figure 4.3). The implemented KNIME-CDK renderer features improved hydrogen layouts and smoother molecule depictions, i.e. rectangular bounding boxes have been replaced by oval bounding boxes, removing ragged corners and displaced bonds. CDK atom colors have been replaced by the popular Rasmol/Chime CPK color scheme (Corey, Pauling, and later Koltun). Differing from CPK colors, carbons and hydrogens are colored black and white because of the default background color used in the KNIME desktop application. The colors are defined as light grey in the CPK color code.

Substructures can be highlighted using a KNIME-CDK specific approach that assigns a pre-defined colour directly on to an atom and bond. This information

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

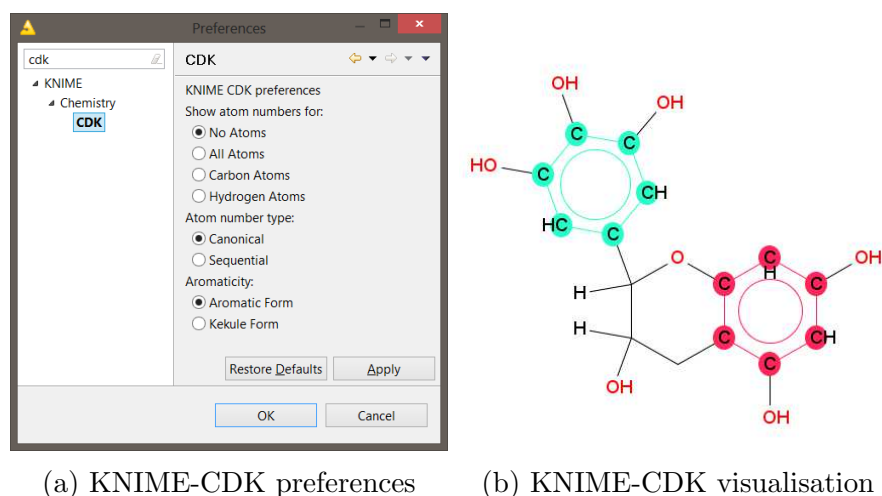


Figure 4.3: Overview and demonstration of KNIME-CDK visualisation preferences. (a) Screenshot of KNIME-CDK’s properties tab. Structures can be drawn in either aromatic or Kekule form with optional atom numbers on all atoms, or carbon/hydrogen atoms only. (b) Example visualisation of Gallocatechin (CHEBI:68330). The benzyl rings are highlighted from an example substructure search using benzole.

is preserved through cycles of disk input and output. It is then globally picked up by the rendering engine on demand.

The KNIME preference page contains a CDK tab to set global visualisation preferences. Given two- or three-dimensional coordinates, compounds are drawn in two-dimensional space with optional atom numbers and in either aromatic or Kekule presentation. Atom numbers can be shown either for all atoms or for Carbon atoms only. Depending on usage, sequential or canonical numbering can be selected. E.g. for HOSE code associations, canonical numbering would be preferred to match atoms to their respective HOSE codes. For visualisation in three dimensions, a separate *3D Viewer* is available based on Jmol.

### 4.4 Evaluation

The KNIME-CDK plug-in was tested using the structurally diverse ChEBI library with a total of 23,240 manually curated structures<sup>[294]</sup>. For testing purposes, the library was used in SDfile format, release 98, because this could arguably



## 4. WORKFLOWS FOR CHEMINFORMATICS

---

be considered the most common use case. For comparison, the well-established RDKit plug-in was used. Using the ChEBI SDfile, consistent input-serialization-output was tested using round tripping. This test ensures that no information is lost or altered.

From the 23,240 structures, 22,225 structures (95.6%) could successfully be read in, marginally less than with the RDKit plug-in (22,482 structures, 96.7%). Not all molecules could be converted into the CDK representation because some classes are not supported throughout the node's read process.

Currently the following groups lack support (examples in brackets, depictions are listed in the appendix section 5): Coordination entities (CHEBI:16304), exotic atoms (CHEBI:27698), complexed porphyrins (CHEBI:27888), some radical species (CHEBI:33101, CHEBI:33105), and repeated structures (CHEBI:65304). The structures were read in  $43.0 \pm 4.5$  seconds compared to  $12.0 \pm 0.7$  seconds (RDKit). Even though the KNIME-CDK plug-in is not as fast as RDKit, which uses a native C++ implementation, its functionality should be seen as complementary to other plug-ins available and its speed is still adequate.

The ChemAxon Marvin Extensions Feature, 2.6.3.v0135, was used to create canonical SMILES from the structures that were loaded with KNIME-CDK and RDKit. For 2794 (12.6%) structures different SMILES were produced, due to the fact that different internal representations and the nature of the problem, inescapably produces variation. This highlights one of the benefits of employing more than one library for processing and analysis tasks. In addition, KNIME-CDK offers some unique functionality including various molecular descriptors, fingerprints, and equivalent class calculation.

For the nightly build of KNIME-CDK, the performance test outlined above was re-run to demonstrate improvements: The ChEBI library SDfile dump, release 112, was used. From the 27780 structures, KNIME-CDK read 27751, 99.9%, successfully (RDKit: 26961, 97.1%). The 29 unsupported structures contain coordination entities (CHEBI:16304), complexed porphyrins (CHEBI:27888), and structures with variable attachment points (CHEBI:51671). The structures were read in  $6.1 \pm 0.9$  seconds (CDK) compared to  $17.4 \pm 0.2$  seconds (RDKit). The

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

nightly build has improved seven-fold to the current standard release. It supports a wider range of chemical classes.

QSPR descriptors are essential for any cheminformatics toolkit. Central descriptors such as the calculated partition coefficient ( $xlogP$ <sup>[295]</sup>) and Lipinski’s Rule of Five<sup>[296]</sup> descriptor have been validated against a set of  $\sim 1,500$  measured octanol water partition coefficients ( $logP$ ) obtained from the TOXNET database<sup>[297]</sup> and a subset of  $\sim 120,000$  molecules from the ZINC database<sup>[298]</sup>, set ‘Clean Drug-Like’ (id:13\_p0.0), respectively.

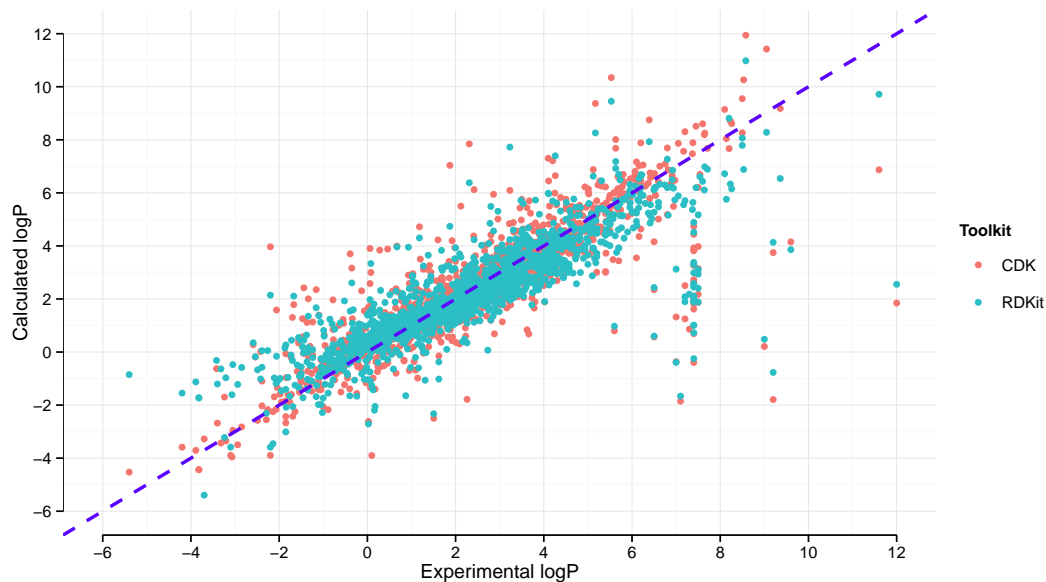
The pairwise comparison of the measured and calculated partition coefficients shows high agreement between the toolkits (Figure 4.4a). For CDK, with sample size  $n = 1589$ , the root-mean-square deviation yields:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^n (\log P - x\log P)^2}{n}} \approx 1.21 \quad (4.1)$$

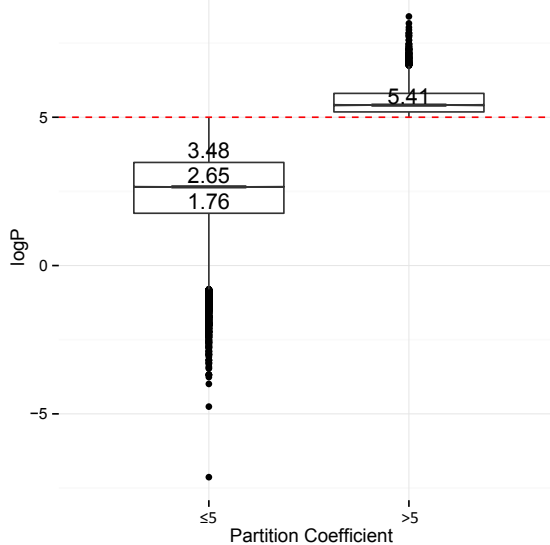
The comparison with RDKit’s RMSD of 1.22 shows a minor but insignificant deviation from CDK’s deviation in calculated partition coefficients. The RMSD represents the standard deviation and summarises the differences between the measured and predicted  $logP$  values. It is a good statistic to indicate the overall performance of CDK’s  $xlogP$ , which works well in KNIME-CDK when compared to RDKit’s reference and the measured  $logP$  values. A RMSD of 1.21 is good for on demand calculated  $logP$  values but it is large when compared to approaches that use experimental information<sup>[299]</sup>.

The Lipinski’s Rule of Five comparison shows few rule violations to the reference set (Table 4.1). A total of 4974 rule violations were detected. 24 violations result from molecules where the rotatable bond count exceeds seven. This was caused by amide groups being incorrectly included in the count. The high rotational barrier of amide C-N bonds disqualify these<sup>[300]</sup>. The descriptor was modified to explicitly ignore amide bonds, resulting in zero rule violations. The remaining 4950 rule violations are caused by ‘incorrectly’ calculated  $logP$  values (Figure 4.4b). Most of the violating values are close to the defined boundary value:  $logP \leq 5$ . The number and magnitude of  $logP$  violations can not solely be explained by

## 4. WORKFLOWS FOR CHEMINFORMATICS



(a) Pairwise comparison of measured and calculated  $\log P$  values



(b) Lipinski's Rule of Five violations

Figure 4.4: Comparison of measured and calculated partition coefficients ( $\log P$ ). (a) Pairwise comparison of partition coefficients calculated by CDK and RDKit to experimentally measured values ( $R_{CDK}^2 = 0.99, R_{RDKit}^2 = 0.99$ ). The 1589 experimental values were obtained from TOXNET. CDK and RDKit give highly similar distributions and deviations (RMSD 1.21 and 1.22). The dashed blue line represents perfect agreement. (b) Boxplots of valid ( $\log P \leq 5$ ) and invalid ( $\log P > 5$ ) descriptor values. As expected, the median of the invalid population ( $n = 4950$ ) is close to the boundary value. Incorrect  $\log P$  values are believed to result from inherent limitations of the algorithm.

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

Rule	Violations
$\log P \leq 5$	4950
<i>rotatable bonds</i> $\leq 7$	0 (24)
<i>H-bond donors</i> $\leq 5$	0
<i>H-bond acceptors</i> $\leq 10$	0
$150 \leq \text{molecular weight} \leq 500$	0

Table 4.1: Table of Lipinski’s Rule of Five violations for the ZINC dataset id:13\_p0.0 (134,322 structures) as calculated by KNIME-CDK. The 24 violations of the rotatable bonds rule result from the inclusion of amide bonds as rotatable bonds. This inclusion is not desirable due to the prohibitively high rotational energy barrier of amide C-N bonds.

an approximate RMSD of 1.2, i.e. deviation by 1.2. The few large outliers – beyond deviation – are believed to result from inherent limits of the algorithm. Inspection of those outliers and cross-comparison with RDKit’s  $\log P$  supports this assumption: both toolkits estimate incorrect values for different subsets. Only 785 molecules have incorrect  $\log P$  values in CDK and RDKit.

### 4.4.1 Round Tripping

Round tripping is a technique to evaluate information loss. In cheminformatics, subtle differences in molecules such as inverted stereochemistry or addition/loss of hydrogen, can significantly change molecular properties. Correct conversion of one chemistry format into another is of paramount importance.

KNIME-CDK was tested using the same ChEBI set from the first round of evaluation (section 4.4). The SDfile was converted to CDK’s molecule representation and then to CML. The CML was back-converted to CDK’s representation and finally to a SDfile. To accommodate for issues during SDfile generation, the SDfile was read in again and converted to CML through CDK’s representation. The resulting chemical representations were then compared to the original SDfile. This process helped to find and eliminate format conversion errors.

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

### 4.4.2 Test Workflows

KNIME-CDK uses workflow tests within the KNIME testing framework for quality assurance. More than a dozen workflows have been created to ensure correct node behaviour in a variety of contexts, including different input formats, KNIME table structures, and execution patterns or structures.

The tests reside on the KNIME testing server and are executed nightly. Different test suits exist following the versioning system outlined in section 4.1. The continuous integration software Jenkins is used as integration tool checking builds and test workflows.

### 4.4.3 Performance and Scalability

The plug-in performs well for tens of thousands of molecules. Whereas most use-cases involve less than millions of molecules, KNIME-CDK has been tested with up to 2,709,359 molecules taken from PubChem. Available memory and the assigned number of threads are the two key limiting factors with regard to execution speed.

An additional comparison with RDKit and Indigo<sup>[301]</sup> shows that the plug-in's bottleneck is molecule conversion (Figure 4.5a). Calculation or query performance on converted molecules are comparable to RDKit and Indigo. Conversion to CML combined with atom typing and aromaticity perception are the primary causes for slow molecule conversion. Other toolkits use the one-line SMILES notation that has high information density. However, SMILES notation does not allow additional molecule information to be stored, a crucial drawback for CDK that relies on atom types and perceived aromaticity. An additional speed penalty comes from the used compression method to reduce disk requirements. Using ChEBI, a seven fold reduction in CML file size is achieved on average during serialization (14 kB to 2 kB) at the cost of seven seconds.

The bottleneck described above has been overcome in the nightly version of the plug-in, which explores the advantages of SMILES: current evidence suggests that

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

SMILES outperforms CML on the basis of speed and resource requirements. However, continuous re-perception of aromaticity and atom types – the advantages of CML – increases the execution time. Increased in-memory execution time appears preferably over disk IO operations for IO-bound frameworks such as KNIME (Figure 4.5b). Improvements in the CDK library in combination with improved KNIME-CDK routines invert the picture seen in the comparison with the stable builds. The nightly build outperforms RDKit and Indigo integrations.

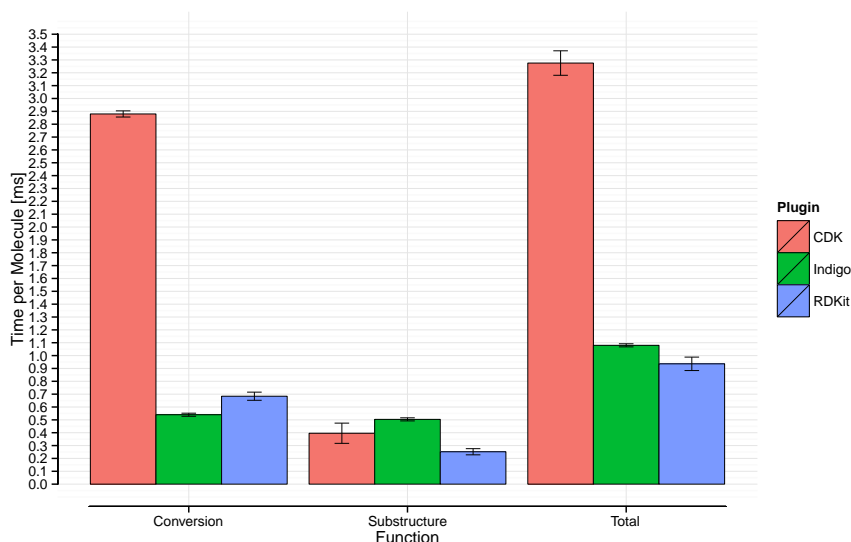
### 4.5 Conclusion

KNIME-CDK is a community-driven plug-in with an active user base. The plug-in is downloaded up to 600 times per month – download statistics are available from the KNIME website, December 2013. It is actively maintained with efforts focussing on extending the implemented functionality and improving overall usability by automating standardization tasks. Over the last couple of years, the underlying methods have been improved dramatically making the plug-in robust and reliable. Together with improvements in the CDK library, a SMILES-backed representation of molecular compounds appears to be the future. Molecule normalization, e.g. tautomer selection and  $pK_a$  calculations, remain dominant challenges to be addressed by the plug-in and the field of cheminformatics. To this end, the tautomer generator will be extended and a set of nodes for normalization will be introduced in the future.

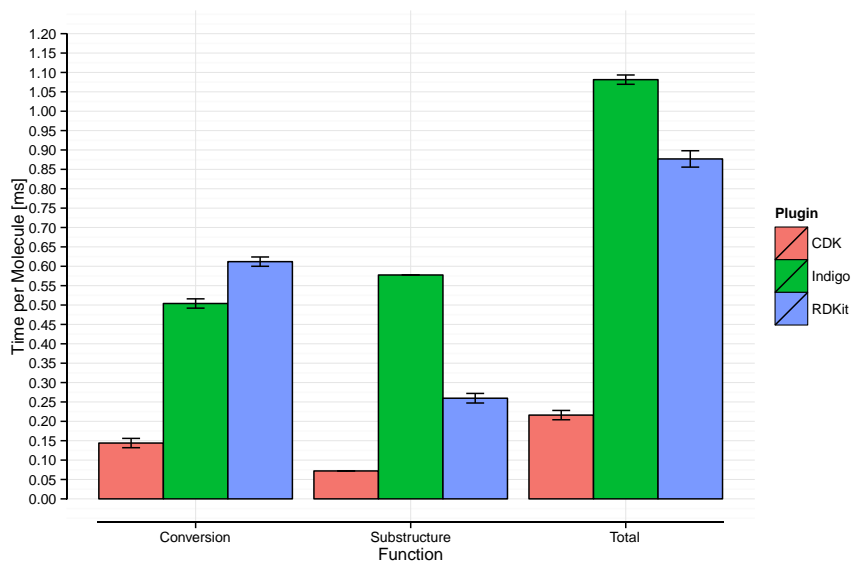
### 4.6 Software Availability

KNIME-CDK and its JChemPaint, AMBIT, and Jmol dependencies have been released under the GNU Lesser General Public License version 3. It depends on the KNIME chemistry types, which are automatically added on installation.

## 4. WORKFLOWS FOR CHEMINFORMATICS



(a) KNIME execution time comparison: v 2.9



(b) KNIME execution time comparison: nightly build

Figure 4.5: Comparison of average execution times per molecule by cheminformatics plug-in. The total time measured equals the sum of the molecule conversion and phenyl substructure search. (a) CDK version 1.4.4, Indigo version 1.1.13, and RDKit version 2.2.0 were used for the comparison. KNIME-CDK is five fold slower than the other plug-ins during conversion and as fast as the other plug-ins during substructure search. Molecule conversion is the bottleneck of KNIME-CDK. (b) Comparison of nightly builds (CDK version 1.5.1, Indigo version 1.1.13, and RDKit version 2.4.0): KNIME-CDK is on average four and a half fold faster than the other plug-ins.

## 4. WORKFLOWS FOR CHEMINFORMATICS

---

### 4.6.1 Update Site

The project is part of KNIME's Community Contributions and hosted on their servers. It can be installed via the official Community Contributions update sites. The entry pages are listed below:

- KNIME-CDK: <https://bitbucket.org/sbeisken/cdkknime>
- KNIME-CDK Support: <http://tech.knime.org/community/cdk>
- KNIME Update Site:  
<http://tech.knime.org/update/community-contributions/<version>>

### 4.6.2 Extensions

KNIME-CDK is a community-driven project that develops by improvements on its core library CDK and feature requests in the KNIME-CDK forum. Node development follows the KNIME node development guidelines. Contributions should be directed to the central KNIME-CDK subversion repository on the KNIME servers.

- KNIME-CDK Forum: <http://tech.knime.org/forum/cdk>

### 4.6.3 Example Workflows

Example workflows illustrating the basics of KNIME-CDK and more advanced applications, such as R integration and data mining, have been deposited on MyExperiment<sup>[221]</sup> under the keyword 'KNIME-CDK'.

Use cases of workflows using KNIME-CDK include the management and analysis of chemical libraries through molecular descriptors, conformer analysis via RMSD, and NMR spectra prediction.

The work has been published in BMC Bioinformatics: Beisken *et al.*: KNIME-CDK: Workflow-driven Cheminformatics. *BMC Bioinformatics* 2013, **14**:257. doi:10.1186/1471-2105-14-257



### SUMMARY AND DISCUSSION

---

Metabolomics is based on analytical technologies that capture data about low molecular weight compounds. Information contained in the data can be distilled through careful data processing and analysis. The high chemical diversity, concentration differences, and dynamics of the metabolome complicate this task.

Detailed electronic logbooks have become fashionable in experimental laboratories. The need to additionally capture the complete digital analysis process as well is becoming a dominant theme in metabolomics. Transparency and reproducibility are key elements to guarantee scalable research within the science community. The ability to share and reuse not only data but also workflows or pipelines is important. Open, ‘copyleft’, licenses form the foundation for data exchange and boundary-free informatics in metabolomics.

Tandem mass spectrometry systems coupled to liquid chromatography (LC-MS<sup>n</sup>) capture detailed metabolomics snapshots of biological samples under controlled conditions. The resulting data sets are convoluted, noisy, and their fine structure is dependent on a plethora of parameters and external factors. Full control over the applied data processing steps is required. Modularity, i.e. the effective addition or removal of individual functions, extendibility, and visual feedback are desirable criteria for any processing tool applied to the task of LC-MS<sup>n</sup> data exploration and analysis.

## 5. SUMMARY AND DISCUSSION

---

*Informatics for LC-MS<sup>n</sup> Analysis:* We created MassCascade, a framework to rapidly analyse and visualise LC-MS<sup>n</sup> data. The framework covers all processing steps from data input to the final generation of the feature matrix. It features a plug-in for the popular workflow platform KNIME, which integrates complementary statistical, bio- and cheminformatics functionality that can be used in combination with MassCascade-KNIME. The plug-in offers a structured approach to data processing in metabolomics that is accessible to bioinformaticians and experimental scientists alike.

Developed in collaboration with the Syngenta AG, MassCascade does not aim to reproduce existing processing tools for LC-MS<sup>n</sup> data but offers a different approach following the paradigm of visual programming. Its focus on feature filtering with multiple ways to remove irrelevant features and its rapid applicability to repetitive processing steps, make it useful for calibration exercises or spectral fingerprinting.

*Knowledge-based Compound Identification:* We added an identification and scoring framework to MassCascade that also integrates seamlessly into the workflow plug-in. The framework enables the generation of reference or spectra libraries for multiple MS level from reference data. It automatically scores and ranks retrieved feature annotations from these reference libraries based on information aggregated during processing, e.g. from isotopic envelopes and from the putative structures themselves. The ranked lists of annotations for every feature of interest simplify metabolite identification.

The metabolites extracted from the study on metabolite ripening demonstrates the use of the methodology and highlights the importance of fragment information. The combination of uni- and multivariate statistics with the ability to rapidly change back to the processing side and fine-tune the workflow using purely open tools, guarantees flexibility and transparency that dramatically facilitates data sharing, e.g. study deposition in the MetaboLights repository or submission of a Nature Scientific Data manuscript.

## 5. SUMMARY AND DISCUSSION

---

*Workflows for Cheminformatics:* LC-MS<sup>n</sup> post-processing relies on cheminformatics for the convenience of handling and visualizing putative metabolites. From the input of small molecule collections, e.g. from plain files, to the selection of molecule subsets based on molecular properties or the generation of molecular formulas from exact molecular masses, cheminformatics has many applications in the context of metabolomics mass spectrometry. We contributed a cheminformatics plug-in based on the Chemistry Development Kit (CDK) to the workflow platform that also hosts MassCascade to provide such functionality and facilitate metabolite identification efforts.

The use cases of the user base exceed far beyond applications in LC-MS<sup>n</sup>. The tool is primarily used as solid cheminformatics toolkit. With a broad user base, it is primarily community driven. Together with our tools and other external plug-ins, it forms a powerful, consolidated set for applications in the life sciences.

Future work involves fine-tuning and further validation of the implemented methods. Differences in resolution of mass spectrometers is only one factor that can have a dramatic impact on data analysis. To provide starting points for data analysis for different instruments, we aim to establish workflows for these instruments that may serve as templates.

To improve the metabolite identification pipeline, we will adopt methods to generate fragmentation spectra (*in silico*) that also give predicted signal abundances. As discussed, this is most likely to increase the power of the scoring framework for annotations from highly similar molecular species.

The KNIME-CDK plug-in evolves continuously together with the CDK core library. To increase its use, molecule normalisation has been singled out as primary area for future work. A collection of nodes will be implemented to simplify the management of sets of molecules.

The high throughput and multivariate data generating nature of metabolomics experiments requires substantial informatics. It the responsibility of the informatics toolkits to simplify data analysis as far as possible while recording everything within the processing pipeline. The tools developed and tested here have been designed to follow these principles. Whereas spectral fingerprinting of biological

## 5. SUMMARY AND DISCUSSION

---

samples is challenging, the major bottleneck in metabolomics is metabolite identification. The aggregation and rationalisation of information from mass spectrometry experiments in a structured manner enables the generation of ranked lists of identifications that also reflect confidence. Identifications with certainty are rare however and require additional evidence only attainable through laboratory work. Tools that facilitate the complete process and show the complexity of processing and ambiguity in identification are small steps in the right direction.

Metabolomics is at a junction. While experimental systems are improving, reproducible metabolite identification is still challenging in metabolomics studies. This is particularly true for unknown metabolites that have not been recorded in previous studies. Whereas some instrumental systems can confidently detect many small molecules from the core metabolism, going beyond that into truly untargeted metabolomics will depend on the existence of high quality, large-scale reference databases and informed information aggregation.

---

# APPENDIX

---

## Univariate Statistics for Feature Analysis

The following figures describe the 13 unique features investigated in chapter 3.5. Univariate statistics for the seven features not discussed in chapter 3.5.3 are shown after. Pairwise Person's correlation matrices were built for each genotype with asymptotic p-values encoded in a three star asterisk representation.

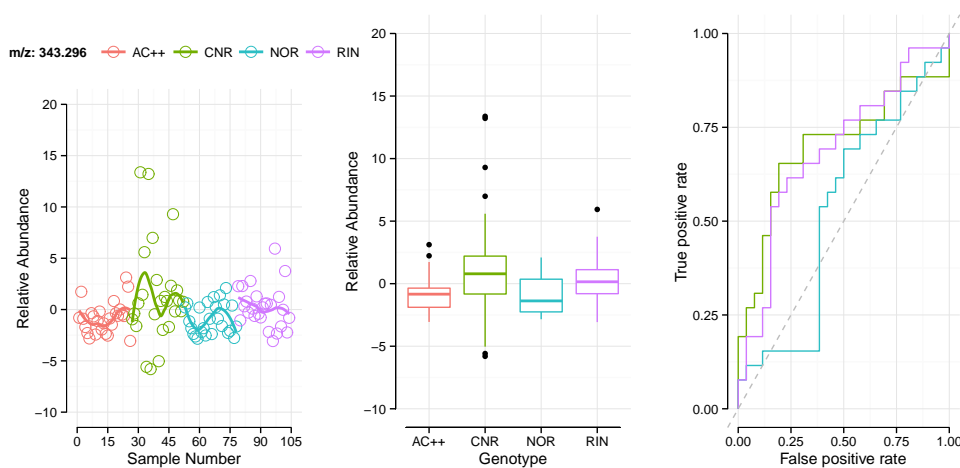
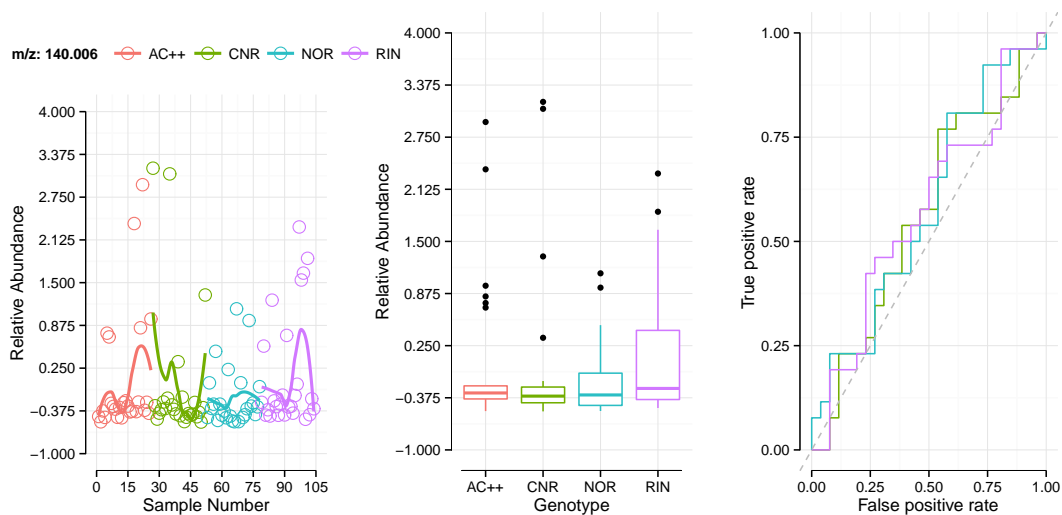
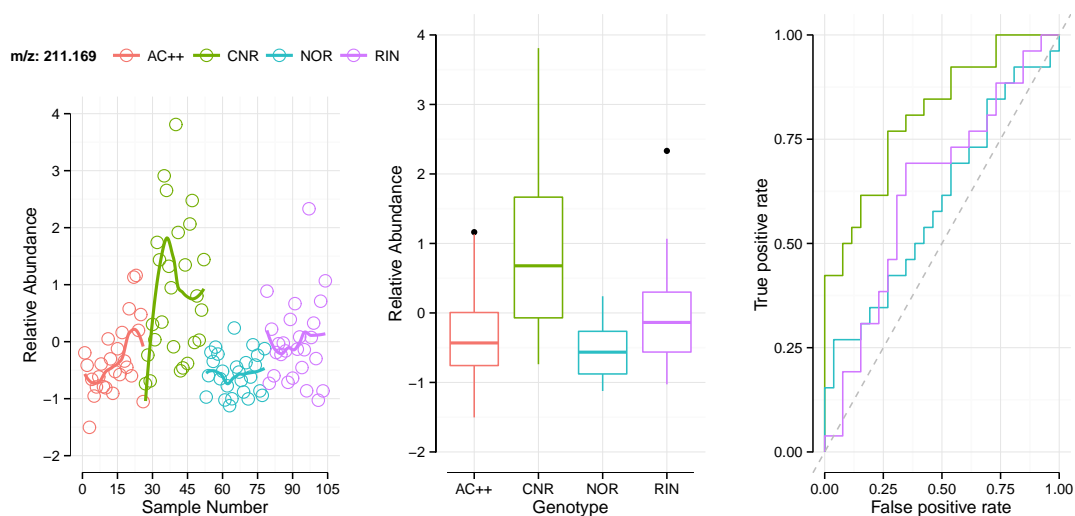


Figure 1: Univariate statistics for feature 343.296. The scatterplot shows feature abundances grouped by genotype and sorted by harvest date in ascending order with applied locally weighted smoothing. The boxplots display the median (middle line) and lower (25%) and upper (75%) percentiles of the abundances. The ROC curves indicate the binary classification power of the mutant genotypes versus the wild type AC<sup>++</sup>.

## Appendix



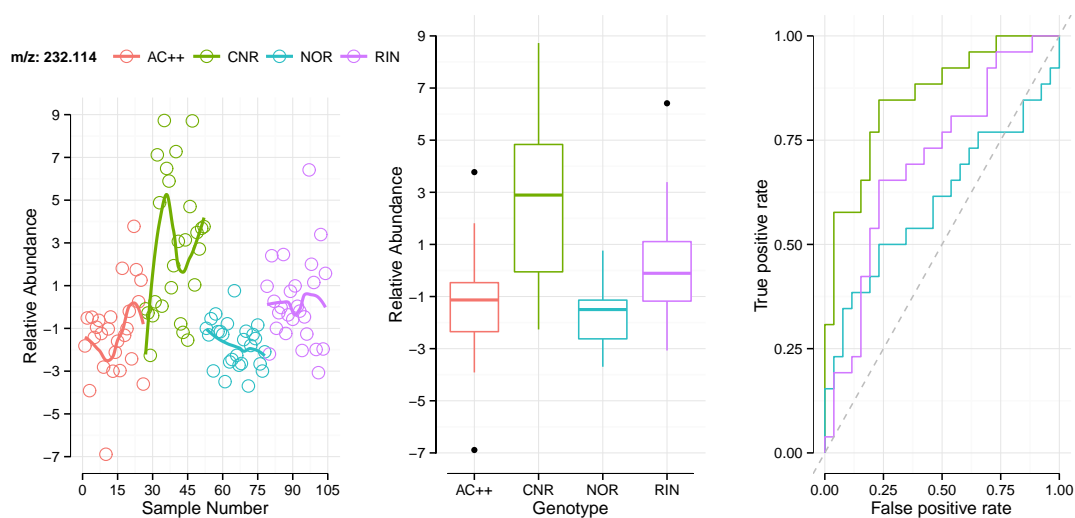
(a) Univariate statistics for feature 140.006



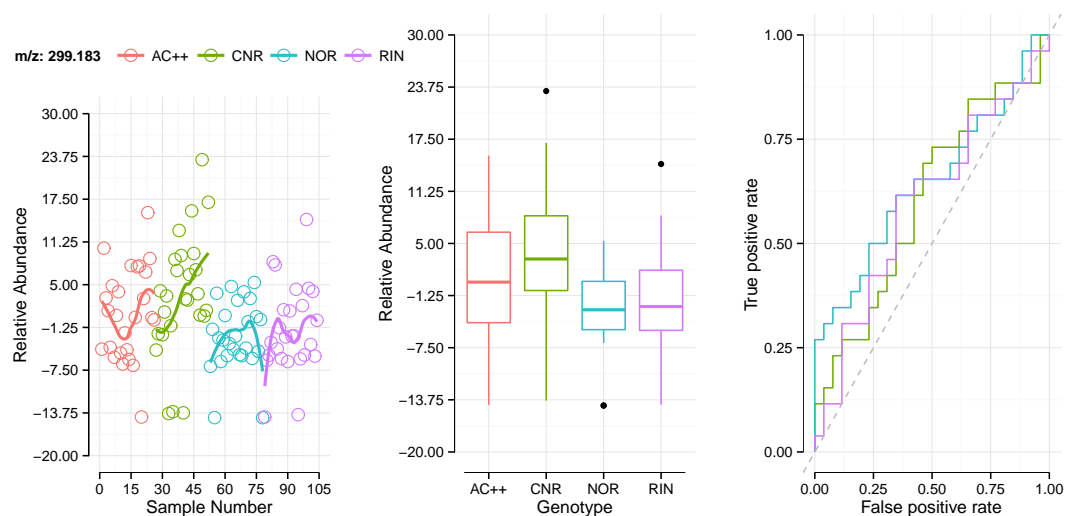
(b) Univariate statistics for feature 211.169

Figure 2: Univariate statistics for features 140.006 (a) and 211.169 (b). The scatterplot shows feature abundances grouped by genotype and sorted by harvest date in ascending order with applied locally weighted smoothing. The boxplots display the median (middle line) and lower (25%) and upper (75%) percentiles of the abundances. The ROC curves indicate the binary classification power of the mutant genotypes versus the wild type AC<sup>++</sup>.

## Appendix



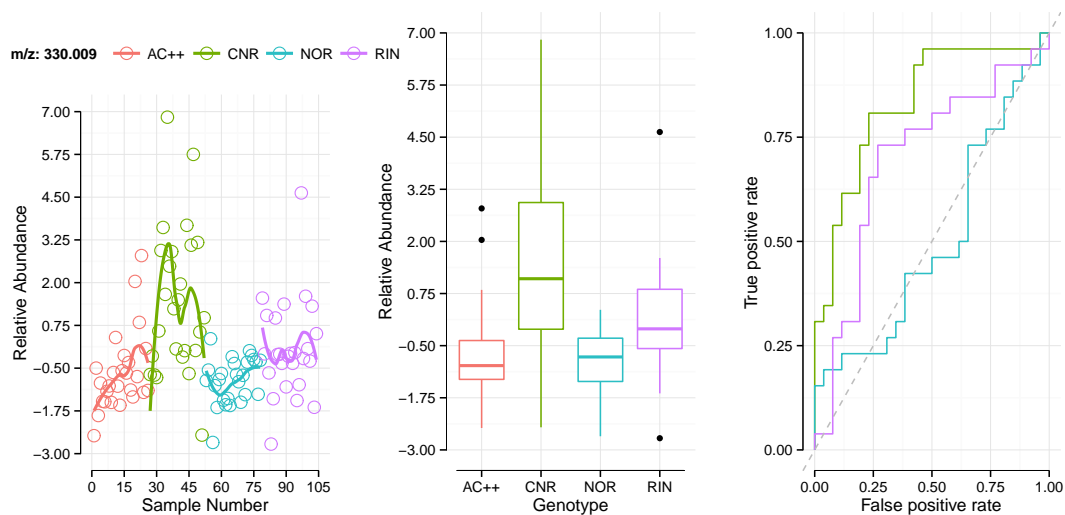
(a) Univariate statistics for feature 232.114



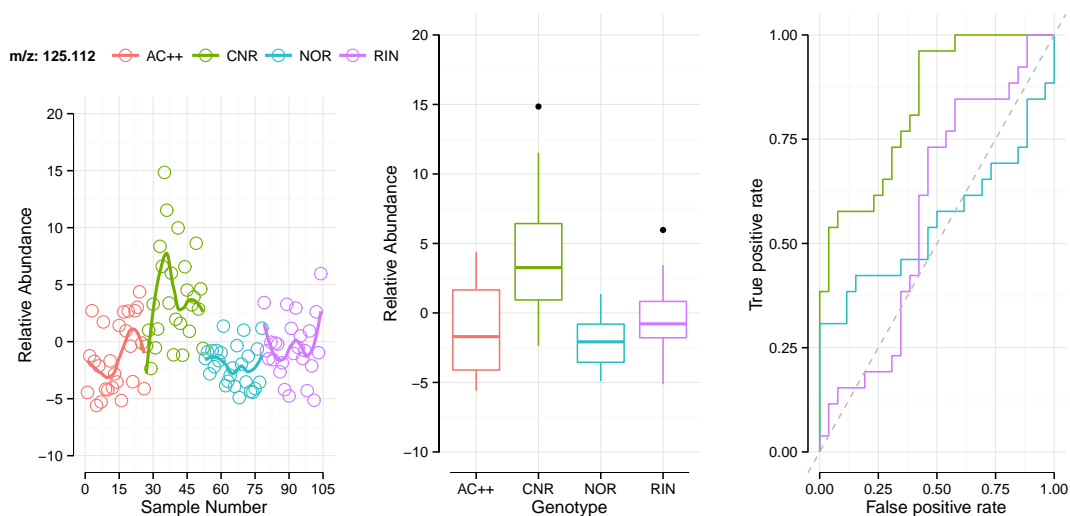
(b) Univariate statistics for feature 299.183

Figure 3: Univariate statistics for features 232.114 (a) and 299.183 (b). The scatterplot shows feature abundances grouped by genotype and sorted by harvest date in ascending order with applied locally weighted smoothing. The boxplots display the median (middle line) and lower (25%) and upper (75%) percentiles of the abundances. The ROC curves indicate the binary classification power of the mutant genotypes versus the wild type AC<sup>++</sup>.

## Appendix



(a) Univariate statistics for feature 330.009

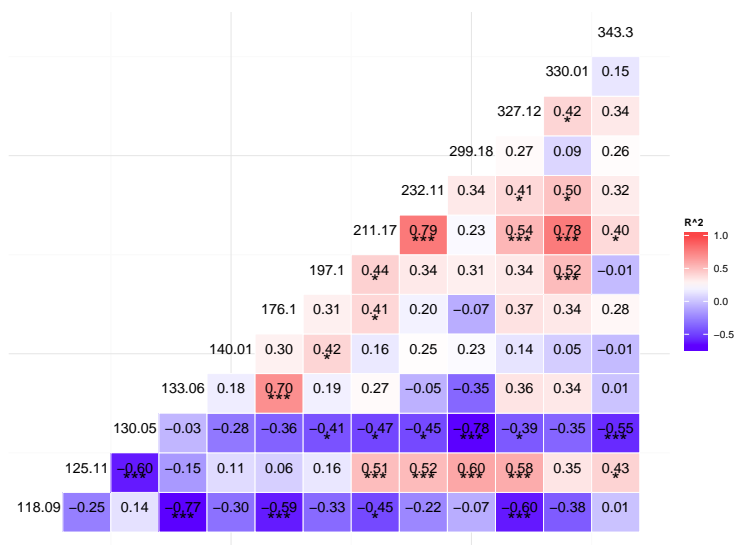


(b) Univariate statistics for feature 125.112

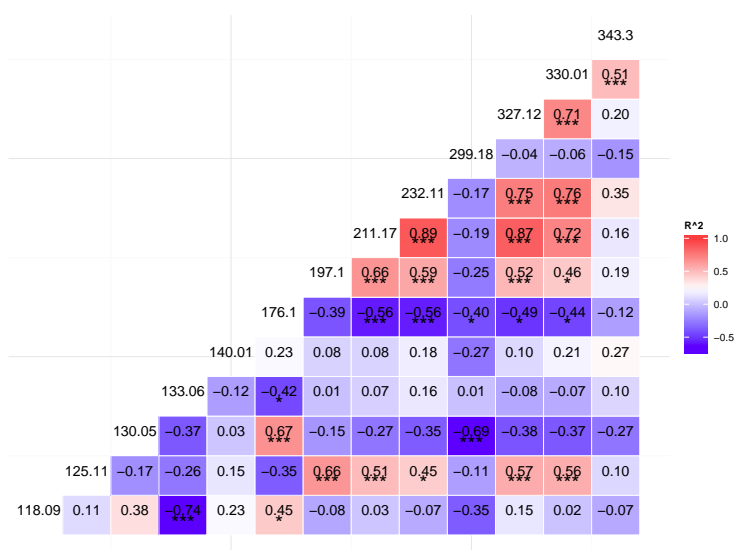
Figure 4: Univariate statistics for features 330.009 (a) and 125.112 (b). The scatterplot shows feature abundances grouped by genotype and sorted by harvest date in ascending order with applied locally weighted smoothing. The boxplots display the median (middle line) and lower (25%) and upper (75%) percentiles of the abundances. The ROC curves indicate the binary classification power of the mutant genotypes versus the wild type AC<sup>++</sup>.



## Appendix



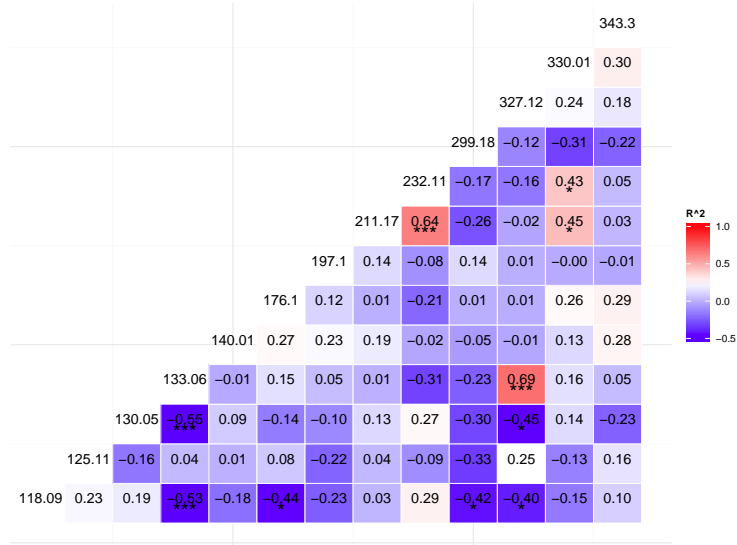
(a) Pairwise Pearson's correlation matrix for WT AC<sup>++</sup>



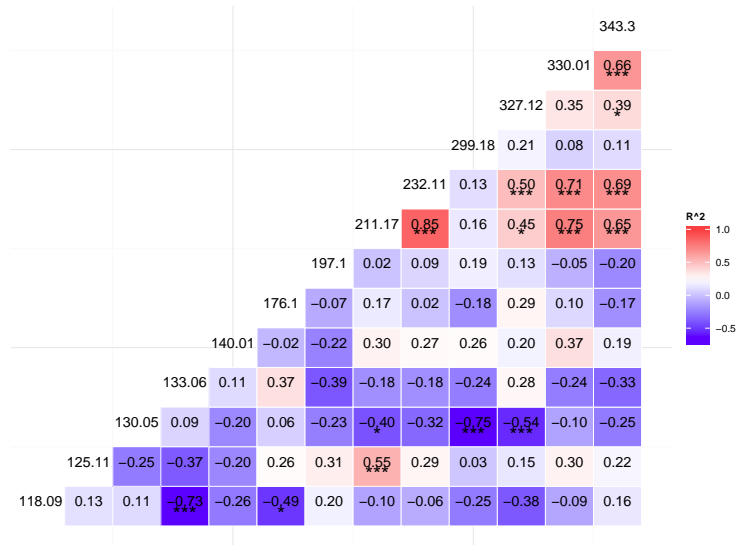
(b) Pairwise Pearson's correlation matrix for mutant CNR

Figure 5: Pearson's pairwise correlation matrix for wild type AC<sup>++</sup> (a) and mutant CNR (b). The asymptotic p-values are encoded using asterisks:  $p \leq 0.01 = ***$ ,  $0.01 < p \leq 0.05 = *$ . The correlation coefficient ranges from low,  $R^2 \leq -0.5$  (dark blue), to high,  $R^2 \geq 0.5$  (bright red).

## Appendix



(a) Pearson's pairwise correlation matrix for mutant NOR



(b) Pearson's pairwise correlation matrix for mutant RIN

Figure 6: Pearson's pairwise correlation matrix for mutants NOR (a) and RIN (b). The asymptotic p-values are encoded using asterisks:  $p \leq 0.01 = ***$ ,  $0.01 < p \leq 0.05 = *$ . The correlation coefficient ranges from low,  $R^2 \leq -0.5$  (dark blue), to high,  $R^2 \geq 0.5$  (bright red).

## Identification

### Rankings

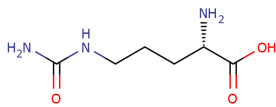
Name	Depiction	Score		
		Frag.	Isotope	None
L-citrulline*		n/a	260	160

Table 1: Metabolite annotation for  $m/z$  176.103 with cumulative scores for no, isotope, and fragment filtering. Missingness and element filters were always included. The calculated score does not allow for selective ranking. All structures are highly similar. \*Matching retention time information in the in-house library at 95 seconds.

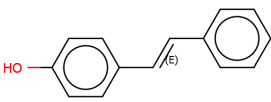
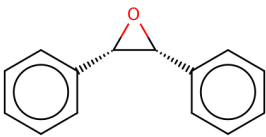
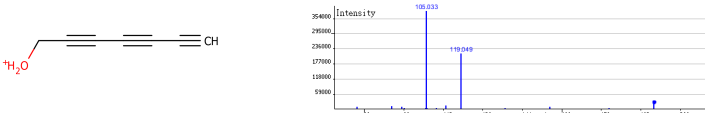
Name	Depiction	Score		
		Frag.	Isotope	None
trans-4-hydroxystilbene <sup>†</sup>		369	251	160
cis-stilbene oxide		369	251	160
MS <sup>2</sup> spectrum				

Table 2: Metabolite annotations for  $m/z$  197.096 with cumulative scores for no, isotope, and fragment filtering. Missingness and element filters were always included. The calculated score does not allow for selective ranking. All structures are highly similar. <sup>†</sup>The extracted MS<sup>2</sup> spectrum shown at the bottom of the table is dominated by a peak at 105.033 (putative structure shown).

## Appendix

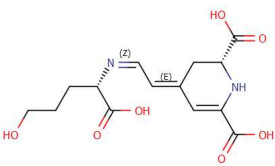
Name	Depiction	Score		
		Frag.	Isotope	None
humilixanthin		n/a	410	160

Table 3: Metabolite annotations for  $m/z$  327.117 with cumulative scores for no, isotope, and fragment filtering. Missingness and element filters were included by default.

## Molecular Formula Generation

Table 4 lists the 13 features that were singled out for identification. Molecular formulas were calculated for the charge corrected features within a 0.05 amu mass tolerance (Table 5). Structures were generated *via* Molgen using the command below. The ‘badlist’ is contained in the program and contains forbidden structures to be discarded when encountered during the deterministic structure generation process.

```
mgen.exe <file_in> -cycles 0-2 -ringsize 0-10  
-o <file_out> -stop 50 -badlist badlist.sdf
```

Feature $m/z$	Feature $m/z$
118.086	197.096
125.112	211.169
130.050	232.114
133.061	299.183
140.011	327.118
176.103	330.009
343.296	

Table 4: Extracted features for identification.

Molecular Weight	Molecular Formula	Molecular Weight	Molecular Formula
117.029	$\text{CH}_3\text{N}_5\text{O}_2$	210.141	$\text{CH}_{14}\text{N}_{12}\text{O}$
117.033	$\text{C}_6\text{H}_3\text{N}_3$	210.153	$\text{H}_{14}\text{N}_{14}$
117.04	$\text{H}_3\text{N}_7\text{O}$	231.056	$\text{CH}_9\text{N}_7\text{O}_7$
117.043	$\text{C}_4\text{H}_7\text{NO}_3$	231.058	$\text{C}_2\text{H}_5\text{N}_{11}\text{O}_3$

## Appendix

---

Molecular Weight	Molecular Formula	Molecular Weight	Molecular Formula
117.054	C <sub>3</sub> H <sub>7</sub> N <sub>3</sub> O <sub>2</sub>	231.059	C <sub>5</sub> H <sub>13</sub> NO <sub>9</sub>
117.058	C <sub>8</sub> H <sub>7</sub> N	231.06	C <sub>6</sub> H <sub>9</sub> N <sub>5</sub> O <sub>5</sub>
117.065	C <sub>2</sub> H <sub>7</sub> N <sub>5</sub> O	231.062	C <sub>7</sub> H <sub>5</sub> N <sub>9</sub> O
117.076	CH <sub>7</sub> N <sub>7</sub>	231.064	C <sub>11</sub> H <sub>9</sub> N <sub>3</sub> O <sub>3</sub>
117.079	C <sub>5</sub> H <sub>11</sub> NO <sub>2</sub>	231.068	H <sub>9</sub> N <sub>9</sub> O <sub>6</sub>
117.09	C <sub>4</sub> H <sub>11</sub> N <sub>3</sub> O	231.069	CH <sub>5</sub> N <sub>13</sub> O <sub>2</sub>
117.101	C <sub>3</sub> H <sub>11</sub> N <sub>5</sub>	231.07	C <sub>4</sub> H <sub>13</sub> N <sub>3</sub> O <sub>8</sub>
117.115	C <sub>6</sub> H <sub>15</sub> NO	231.072	C <sub>5</sub> H <sub>9</sub> N <sub>7</sub> O <sub>4</sub>
117.127	C <sub>5</sub> H <sub>15</sub> N <sub>3</sub>	231.073	C <sub>6</sub> H <sub>5</sub> N <sub>11</sub>
124.06	CH <sub>8</sub> N <sub>4</sub> O <sub>3</sub>	231.074	C <sub>9</sub> H <sub>13</sub> NO <sub>6</sub>
124.064	C <sub>6</sub> H <sub>8</sub> N <sub>2</sub> O	231.076	C <sub>10</sub> H <sub>9</sub> N <sub>5</sub> O <sub>2</sub>
124.071	H <sub>8</sub> N <sub>6</sub> O <sub>2</sub>	231.08	C <sub>15</sub> H <sub>9</sub> N <sub>3</sub>
124.075	C <sub>5</sub> H <sub>8</sub> N <sub>4</sub>	231.08	H <sub>5</sub> N <sub>15</sub> O
124.089	C <sub>8</sub> H <sub>12</sub> O	231.081	C <sub>3</sub> H <sub>13</sub> N <sub>5</sub> O <sub>7</sub>
124.1	C <sub>7</sub> H <sub>12</sub> N <sub>2</sub>	231.083	C <sub>4</sub> H <sub>9</sub> N <sub>9</sub> O <sub>3</sub>
129.006	C <sub>4</sub> H <sub>3</sub> NO <sub>4</sub>	231.086	C <sub>8</sub> H <sub>13</sub> N <sub>3</sub> O <sub>5</sub>
129.017	C <sub>3</sub> H <sub>3</sub> N <sub>3</sub> O <sub>3</sub>	231.087	C <sub>9</sub> H <sub>9</sub> N <sub>7</sub> O
129.021	C <sub>8</sub> H <sub>3</sub> NO	231.09	C <sub>13</sub> H <sub>13</sub> NO <sub>3</sub>
129.029	C <sub>2</sub> H <sub>3</sub> N <sub>5</sub> O <sub>2</sub>	231.093	C <sub>2</sub> H <sub>13</sub> N <sub>7</sub> O <sub>6</sub>
129.033	C <sub>7</sub> H <sub>3</sub> N <sub>3</sub>	231.094	C <sub>3</sub> H <sub>9</sub> N <sub>11</sub> O <sub>2</sub>
129.04	CH <sub>3</sub> N <sub>7</sub> O	231.097	C <sub>7</sub> H <sub>13</sub> N <sub>5</sub> O <sub>4</sub>
129.043	C <sub>5</sub> H <sub>7</sub> NO <sub>3</sub>	231.098	C <sub>8</sub> H <sub>9</sub> N <sub>9</sub>
129.051	H <sub>3</sub> N <sub>9</sub>	231.101	C <sub>12</sub> H <sub>13</sub> N <sub>3</sub> O <sub>2</sub>
129.054	C <sub>4</sub> H <sub>7</sub> N <sub>3</sub> O <sub>2</sub>	231.104	CH <sub>13</sub> N <sub>9</sub> O <sub>5</sub>
129.058	C <sub>9</sub> H <sub>7</sub> N	231.105	C <sub>2</sub> H <sub>9</sub> N <sub>13</sub> O
129.065	C <sub>3</sub> H <sub>7</sub> N <sub>5</sub> O	231.108	C <sub>6</sub> H <sub>13</sub> N <sub>7</sub> O <sub>3</sub>
129.076	C <sub>2</sub> H <sub>7</sub> N <sub>7</sub>	231.112	C <sub>11</sub> H <sub>13</sub> N <sub>5</sub> O
129.079	C <sub>6</sub> H <sub>11</sub> NO <sub>2</sub>	231.115	H <sub>13</sub> N <sub>11</sub> O <sub>4</sub>
129.09	C <sub>5</sub> H <sub>11</sub> N <sub>3</sub> O	231.117	CH <sub>9</sub> N <sub>15</sub>
132.003	N <sub>6</sub> O <sub>3</sub>	231.119	C <sub>5</sub> H <sub>13</sub> N <sub>9</sub> O <sub>2</sub>
132.006	C <sub>4</sub> H <sub>4</sub> O <sub>5</sub>	231.123	C <sub>10</sub> H <sub>13</sub> N <sub>7</sub>
132.007	C <sub>5</sub> N <sub>4</sub> O	231.13	C <sub>4</sub> H <sub>13</sub> N <sub>11</sub> O
132.017	C <sub>3</sub> H <sub>4</sub> N <sub>2</sub> O <sub>4</sub>	231.142	C <sub>3</sub> H <sub>13</sub> N <sub>13</sub>
132.018	C <sub>4</sub> N <sub>6</sub>	298.129	C <sub>13</sub> H <sub>14</sub> N <sub>8</sub> O
132.021	C <sub>8</sub> H <sub>4</sub> O <sub>2</sub>	298.132	C <sub>2</sub> H <sub>14</sub> N <sub>14</sub> O <sub>4</sub>
132.028	C <sub>2</sub> H <sub>4</sub> N <sub>4</sub> O <sub>3</sub>	298.136	C <sub>7</sub> H <sub>14</sub> N <sub>12</sub> O <sub>2</sub>
132.032	C <sub>7</sub> H <sub>4</sub> N <sub>2</sub> O	298.14	C <sub>12</sub> H <sub>14</sub> N <sub>10</sub>
132.04	CH <sub>4</sub> N <sub>6</sub> O <sub>2</sub>	298.148	C <sub>6</sub> H <sub>14</sub> N <sub>14</sub> O
132.042	C <sub>5</sub> H <sub>8</sub> O <sub>4</sub>	326.061	C <sub>10</sub> H <sub>10</sub> N <sub>6</sub> O <sub>7</sub>
132.044	C <sub>6</sub> H <sub>4</sub> N <sub>4</sub>	326.062	C <sub>11</sub> H <sub>6</sub> N <sub>10</sub> O <sub>3</sub>

## Appendix

---

Molecular Weight	Molecular Formula	Molecular Weight	Molecular Formula
132.051	H <sub>4</sub> N <sub>8</sub> O	326.064	C <sub>14</sub> H <sub>14</sub> O <sub>9</sub>
132.053	C <sub>4</sub> H <sub>8</sub> N <sub>2</sub> O <sub>3</sub>	326.065	C <sub>15</sub> H <sub>10</sub> N <sub>4</sub> O <sub>5</sub>
132.058	C <sub>9</sub> H <sub>8</sub> O	326.067	C <sub>3</sub> H <sub>14</sub> N <sub>6</sub> O <sub>12</sub>
132.065	C <sub>3</sub> H <sub>8</sub> N <sub>4</sub> O <sub>2</sub>	326.068	C <sub>4</sub> H <sub>10</sub> N <sub>10</sub> O <sub>8</sub>
132.069	C <sub>8</sub> H <sub>8</sub> N <sub>2</sub>	326.07	C <sub>5</sub> H <sub>6</sub> N <sub>14</sub> O <sub>4</sub>
132.076	C <sub>2</sub> H <sub>8</sub> N <sub>6</sub> O	326.071	C <sub>8</sub> H <sub>14</sub> N <sub>4</sub> O <sub>10</sub>
132.079	C <sub>6</sub> H <sub>12</sub> O <sub>3</sub>	326.072	C <sub>9</sub> H <sub>10</sub> N <sub>8</sub> O <sub>6</sub>
132.087	CH <sub>8</sub> N <sub>8</sub>	326.074	C <sub>10</sub> H <sub>6</sub> N <sub>12</sub> O <sub>2</sub>
132.09	C <sub>5</sub> H <sub>12</sub> N <sub>2</sub> O <sub>2</sub>	326.075	C <sub>13</sub> H <sub>14</sub> N <sub>2</sub> O <sub>8</sub>
132.094	C <sub>10</sub> H <sub>12</sub>	326.076	C <sub>14</sub> H <sub>10</sub> N <sub>6</sub> O <sub>4</sub>
132.101	C <sub>4</sub> H <sub>12</sub> N <sub>4</sub> O	326.078	C <sub>15</sub> H <sub>6</sub> N <sub>10</sub>
138.975	CHNO <sub>7</sub>	326.078	C <sub>2</sub> H <sub>14</sub> N <sub>8</sub> O <sub>11</sub>
138.987	HN <sub>3</sub> O <sub>6</sub>	326.08	C <sub>3</sub> H <sub>10</sub> N <sub>12</sub> O <sub>7</sub>
138.991	C <sub>5</sub> HNO <sub>4</sub>	326.082	C <sub>7</sub> H <sub>14</sub> N <sub>6</sub> O <sub>9</sub>
139.002	C <sub>4</sub> HN <sub>3</sub> O <sub>3</sub>	326.084	C <sub>8</sub> H <sub>10</sub> N <sub>10</sub> O <sub>5</sub>
139.006	C <sub>9</sub> HNO	326.085	C <sub>9</sub> H <sub>6</sub> N <sub>14</sub> O
139.012	C <sub>2</sub> H <sub>5</sub> NO <sub>6</sub>	326.086	C <sub>12</sub> H <sub>14</sub> N <sub>4</sub> O <sub>7</sub>
139.013	C <sub>3</sub> HN <sub>5</sub> O <sub>2</sub>	326.088	C <sub>13</sub> H <sub>10</sub> N <sub>8</sub> O <sub>3</sub>
139.017	C <sub>8</sub> HN <sub>3</sub>	326.089	CH <sub>14</sub> N <sub>10</sub> O <sub>10</sub>
139.023	CH <sub>5</sub> N <sub>3</sub> O <sub>5</sub>	326.091	C <sub>2</sub> H <sub>10</sub> N <sub>14</sub> O <sub>6</sub>
139.024	C <sub>2</sub> HN <sub>7</sub> O	326.093	C <sub>6</sub> H <sub>14</sub> N <sub>8</sub> O <sub>8</sub>
139.027	C <sub>6</sub> H <sub>5</sub> NO <sub>3</sub>	326.095	C <sub>7</sub> H <sub>10</sub> N <sub>12</sub> O <sub>4</sub>
139.034	H <sub>5</sub> N <sub>5</sub> O <sub>4</sub>	326.097	C <sub>11</sub> H <sub>14</sub> N <sub>6</sub> O <sub>6</sub>
139.035	CHN <sub>9</sub>	326.099	C <sub>12</sub> H <sub>10</sub> N <sub>10</sub> O <sub>2</sub>
139.038	C <sub>5</sub> H <sub>5</sub> N <sub>3</sub> O <sub>2</sub>	326.101	H <sub>14</sub> N <sub>12</sub> O <sub>9</sub>
139.042	C <sub>10</sub> H <sub>5</sub> N	326.105	C <sub>5</sub> H <sub>14</sub> N <sub>10</sub> O <sub>7</sub>
139.048	C <sub>3</sub> H <sub>9</sub> NO <sub>5</sub>	326.106	C <sub>6</sub> H <sub>10</sub> N <sub>14</sub> O <sub>3</sub>
139.049	C <sub>4</sub> H <sub>5</sub> N <sub>5</sub> O	326.109	C <sub>10</sub> H <sub>14</sub> N <sub>8</sub> O <sub>5</sub>
175.045	C <sub>2</sub> H <sub>5</sub> N <sub>7</sub> O <sub>3</sub>	326.11	C <sub>11</sub> H <sub>10</sub> N <sub>12</sub> O
175.048	C <sub>6</sub> H <sub>9</sub> NO <sub>5</sub>	326.113	C <sub>15</sub> H <sub>14</sub> N <sub>6</sub> O <sub>3</sub>
175.049	C <sub>7</sub> H <sub>5</sub> N <sub>5</sub> O	326.116	C <sub>4</sub> H <sub>14</sub> N <sub>12</sub> O <sub>6</sub>
175.057	CH <sub>5</sub> N <sub>9</sub> O <sub>2</sub>	326.12	C <sub>9</sub> H <sub>14</sub> N <sub>10</sub> O <sub>4</sub>
175.059	C <sub>5</sub> H <sub>9</sub> N <sub>3</sub> O <sub>4</sub>	326.121	C <sub>10</sub> H <sub>10</sub> N <sub>14</sub>
175.061	C <sub>6</sub> H <sub>5</sub> N <sub>7</sub>	326.124	C <sub>14</sub> H <sub>14</sub> N <sub>8</sub> O <sub>2</sub>
175.063	C <sub>10</sub> H <sub>9</sub> NO <sub>2</sub>	326.127	C <sub>3</sub> H <sub>14</sub> N <sub>14</sub> O <sub>5</sub>
175.068	H <sub>5</sub> N <sub>11</sub> O	326.131	C <sub>8</sub> H <sub>14</sub> N <sub>12</sub> O <sub>3</sub>
175.071	C <sub>4</sub> H <sub>9</sub> N <sub>5</sub> O <sub>3</sub>	326.135	C <sub>13</sub> H <sub>14</sub> N <sub>10</sub> O
175.075	C <sub>9</sub> H <sub>9</sub> N <sub>3</sub> O	326.142	C <sub>7</sub> H <sub>14</sub> N <sub>14</sub> O <sub>2</sub>
175.082	C <sub>3</sub> H <sub>9</sub> N <sub>7</sub> O <sub>2</sub>	326.146	C <sub>12</sub> H <sub>14</sub> N <sub>12</sub>
175.084	C <sub>7</sub> H <sub>13</sub> NO <sub>4</sub>	328.962	C <sub>5</sub> H <sub>3</sub> N <sub>3</sub> O <sub>14</sub>

## Appendix

Molecular Weight	Molecular Formula	Molecular Weight	Molecular Formula
175.086	C <sub>8</sub> H <sub>9</sub> N <sub>5</sub>	328.966	C <sub>10</sub> H <sub>3</sub> NO <sub>12</sub>
175.093	C <sub>2</sub> H <sub>9</sub> N <sub>9</sub> O	328.973	C <sub>4</sub> H <sub>3</sub> N <sub>5</sub> O <sub>13</sub>
175.096	C <sub>6</sub> H <sub>13</sub> N <sub>3</sub> O <sub>3</sub>	328.977	C <sub>9</sub> H <sub>3</sub> N <sub>3</sub> O <sub>11</sub>
175.1	C <sub>11</sub> H <sub>13</sub> NO	328.981	C <sub>14</sub> H <sub>3</sub> NO <sub>9</sub>
175.104	CH <sub>9</sub> N <sub>11</sub>	328.984	C <sub>3</sub> H <sub>3</sub> N <sub>7</sub> O <sub>12</sub>
175.107	C <sub>5</sub> H <sub>13</sub> N <sub>5</sub> O <sub>2</sub>	328.987	C <sub>7</sub> H <sub>7</sub> NO <sub>14</sub>
175.111	C <sub>10</sub> H <sub>13</sub> N <sub>3</sub>	328.988	C <sub>8</sub> H <sub>3</sub> N <sub>5</sub> O <sub>10</sub>
175.118	C <sub>4</sub> H <sub>13</sub> N <sub>7</sub> O	328.992	C <sub>13</sub> H <sub>3</sub> N <sub>3</sub> O <sub>8</sub>
175.129	C <sub>3</sub> H <sub>13</sub> N <sub>9</sub>	328.994	CH <sub>7</sub> N <sub>5</sub> O <sub>15</sub>
196.039	C <sub>10</sub> H <sub>4</sub> N <sub>4</sub> O	328.995	C <sub>2</sub> H <sub>3</sub> N <sub>9</sub> O <sub>11</sub>
196.043	N <sub>14</sub>	328.998	C <sub>6</sub> H <sub>7</sub> N <sub>3</sub> O <sub>13</sub>
196.044	C <sub>3</sub> H <sub>8</sub> N <sub>4</sub> O <sub>6</sub>	328.999	C <sub>7</sub> H <sub>3</sub> N <sub>7</sub> O <sub>9</sub>
196.046	C <sub>4</sub> H <sub>4</sub> N <sub>8</sub> O <sub>2</sub>	329.002	C <sub>11</sub> H <sub>7</sub> NO <sub>11</sub>
196.048	C <sub>8</sub> H <sub>8</sub> N <sub>2</sub> O <sub>4</sub>	329.003	C <sub>12</sub> H <sub>3</sub> N <sub>5</sub> O <sub>7</sub>
196.05	C <sub>9</sub> H <sub>4</sub> N <sub>6</sub>	329.005	H <sub>7</sub> N <sub>7</sub> O <sub>14</sub>
196.052	C <sub>13</sub> H <sub>8</sub> O <sub>2</sub>	329.006	CH <sub>3</sub> N <sub>11</sub> O <sub>10</sub>
196.056	C <sub>2</sub> H <sub>8</sub> N <sub>6</sub> O <sub>5</sub>	329.009	C <sub>5</sub> H <sub>7</sub> N <sub>5</sub> O <sub>12</sub>
196.057	C <sub>3</sub> H <sub>4</sub> N <sub>10</sub> O	329.01	C <sub>6</sub> H <sub>3</sub> N <sub>9</sub> O <sub>8</sub>
196.058	C <sub>6</sub> H <sub>12</sub> O <sub>7</sub>	329.013	C <sub>10</sub> H <sub>7</sub> N <sub>3</sub> O <sub>10</sub>
196.06	C <sub>7</sub> H <sub>8</sub> N <sub>4</sub> O <sub>3</sub>	329.014	C <sub>11</sub> H <sub>3</sub> N <sub>7</sub> O <sub>6</sub>
196.064	C <sub>12</sub> H <sub>8</sub> N <sub>2</sub> O	329.017	C <sub>15</sub> H <sub>7</sub> NO <sub>8</sub>
196.067	CH <sub>8</sub> N <sub>8</sub> O <sub>4</sub>	329.018	H <sub>3</sub> N <sub>13</sub> O <sub>9</sub>
196.068	C <sub>2</sub> H <sub>4</sub> N <sub>12</sub>	329.019	C <sub>3</sub> H <sub>11</sub> N <sub>3</sub> O <sub>15</sub>
196.07	C <sub>5</sub> H <sub>12</sub> N <sub>2</sub> O <sub>6</sub>	329.02	C <sub>4</sub> H <sub>7</sub> N <sub>7</sub> O <sub>11</sub>
196.071	C <sub>6</sub> H <sub>8</sub> N <sub>6</sub> O <sub>2</sub>	329.022	C <sub>5</sub> H <sub>3</sub> N <sub>11</sub> O <sub>7</sub>
196.074	C <sub>10</sub> H <sub>12</sub> O <sub>4</sub>	329.023	C <sub>8</sub> H <sub>11</sub> NO <sub>13</sub>
196.075	C <sub>11</sub> H <sub>8</sub> N <sub>4</sub>	329.024	C <sub>9</sub> H <sub>7</sub> N <sub>5</sub> O <sub>9</sub>
196.078	H <sub>8</sub> N <sub>10</sub> O <sub>3</sub>	329.026	C <sub>10</sub> H <sub>3</sub> N <sub>9</sub> O <sub>5</sub>
196.081	C <sub>4</sub> H <sub>12</sub> N <sub>4</sub> O <sub>5</sub>	329.028	C <sub>14</sub> H <sub>7</sub> N <sub>3</sub> O <sub>7</sub>
196.082	C <sub>5</sub> H <sub>8</sub> N <sub>8</sub> O	329.03	C <sub>15</sub> H <sub>3</sub> N <sub>7</sub> O <sub>3</sub>
196.085	C <sub>9</sub> H <sub>12</sub> N <sub>2</sub> O <sub>3</sub>	329.03	C <sub>2</sub> H <sub>11</sub> N <sub>5</sub> O <sub>14</sub>
196.089	C <sub>14</sub> H <sub>12</sub> O	329.032	C <sub>3</sub> H <sub>7</sub> N <sub>9</sub> O <sub>10</sub>
196.092	C <sub>3</sub> H <sub>12</sub> N <sub>6</sub> O <sub>4</sub>	329.033	C <sub>4</sub> H <sub>3</sub> N <sub>13</sub> O <sub>6</sub>
196.093	C <sub>4</sub> H <sub>8</sub> N <sub>10</sub>	329.034	C <sub>7</sub> H <sub>11</sub> N <sub>3</sub> O <sub>12</sub>
196.096	C <sub>8</sub> H <sub>12</sub> N <sub>4</sub> O <sub>2</sub>	329.036	C <sub>8</sub> H <sub>7</sub> N <sub>7</sub> O <sub>8</sub>
196.1	C <sub>13</sub> H <sub>12</sub> N <sub>2</sub>	329.037	C <sub>9</sub> H <sub>3</sub> N <sub>11</sub> O <sub>4</sub>
196.103	C <sub>2</sub> H <sub>12</sub> N <sub>8</sub> O <sub>3</sub>	329.038	C <sub>12</sub> H <sub>11</sub> NO <sub>10</sub>
196.107	C <sub>7</sub> H <sub>12</sub> N <sub>6</sub> O	329.04	C <sub>13</sub> H <sub>7</sub> N <sub>5</sub> O <sub>6</sub>
196.114	CH <sub>12</sub> N <sub>10</sub> O <sub>2</sub>	329.041	C <sub>14</sub> H <sub>3</sub> N <sub>9</sub> O <sub>2</sub>
196.118	C <sub>6</sub> H <sub>12</sub> N <sub>8</sub>	329.041	CH <sub>11</sub> N <sub>7</sub> O <sub>13</sub>

## Appendix

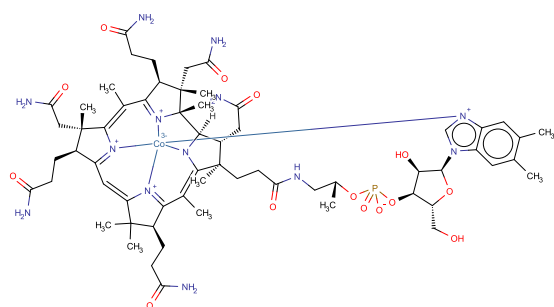
---

Molecular Weight	Molecular Formula	Molecular Weight	Molecular Formula
196.126	H <sub>12</sub> N <sub>12</sub> O	329.043	C <sub>2</sub> H <sub>7</sub> N <sub>11</sub> O <sub>9</sub>
210.112	C <sub>9</sub> H <sub>14</sub> N <sub>4</sub> O <sub>2</sub>	329.044	C <sub>3</sub> H <sub>3</sub> N <sub>15</sub> O <sub>5</sub>
210.116	C <sub>14</sub> H <sub>14</sub> N <sub>2</sub>	329.046	C <sub>6</sub> H <sub>11</sub> N <sub>5</sub> O <sub>11</sub>
210.119	C <sub>3</sub> H <sub>14</sub> N <sub>8</sub> O <sub>3</sub>	329.047	C <sub>7</sub> H <sub>7</sub> N <sub>9</sub> O <sub>7</sub>
210.123	C <sub>8</sub> H <sub>14</sub> N <sub>6</sub> O	329.048	C <sub>8</sub> H <sub>3</sub> N <sub>13</sub> O <sub>3</sub>
210.13	C <sub>2</sub> H <sub>14</sub> N <sub>10</sub> O <sub>2</sub>	329.05	C <sub>11</sub> H <sub>11</sub> N <sub>3</sub> O <sub>9</sub>
210.134	C <sub>7</sub> H <sub>14</sub> N <sub>8</sub>		

Table 5: Molecular formulas calculated from the extracted features within a 0.05 amu mass window. Incorrect formulas were filtered out using the seven golden rules described by Kind *et al.* [\[240\]](#).



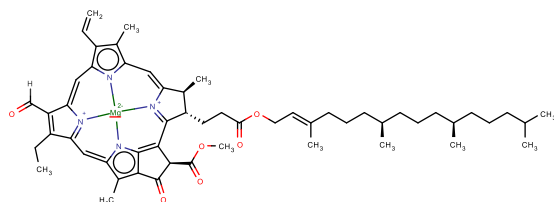
## Unsupported Structures in KNIME-CDK



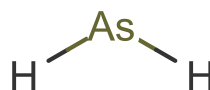
(a) Cob(II)alamin (CHEBI:16304)



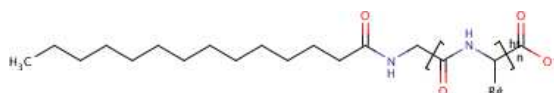
(b) Nitrogen dioxide (CHEBI:33101)



(c) Chlorophyll b (CHEBI:2788)



(d) Arsanyl (CHEBI:33105)



(e) 14-C-Glycylpeptide (CHEBI:65304)



(f) Vanadium (CHEBI:27698)

Figure 7: Examples of chemical classes that lack support in KNIME-CDK. (a) Coordination entities (CHEBI:16304), (b, d) some radical species (CHEBI:33101, CHEBI:33105), (c) complexed porphyrins (CHEBI:27888), (e) repeated structures (CHEBI:65304), and (f) exotic atoms (CHEBI:27698).

## Appendix

### LC-MS Interferents

List of potential interference ions in positive ion mode LC-ESI-MS up to 1000 Da. The list was adapted from Keller *et al.*<sup>[110]</sup>. The ion mass, ion type, formula, and chemical species is provided.

Ion Mass	Ion Type	Formula or Subunit	Species
33.03349	[M+H] <sup>+</sup>	CH <sub>3</sub> OH	Methanol
42.03383	[M+H] <sup>+</sup>	CH <sub>3</sub> CN	ACN
59.06037	[M+NH <sub>4</sub> ] <sup>+</sup>	CH <sub>3</sub> CN	ACN
63.04406	[A <sub>1</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
64.01577	[M+Na] <sup>+</sup>	CH <sub>3</sub> CN	ACN
65.05971	[M <sub>2</sub> +H] <sup>+</sup>	CH <sub>3</sub> OH	Methanol
74.06004	[M+H] <sup>+</sup>	C <sub>3</sub> H <sub>7</sub> NO	Dimethyl formamide
74.06004	[A <sub>1</sub> B <sub>1</sub> +H] <sup>+</sup>	(CH <sub>3</sub> CN) <sub>n</sub> (CH <sub>3</sub> OH) <sub>m</sub>	Acetonitrile/Methanol
77.05971	[A <sub>1</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
79.02121	[M+H] <sup>+</sup>	C <sub>2</sub> H <sub>6</sub> OS	DMSO
83.06037	[M <sub>2</sub> +H] <sup>+</sup>	CH <sub>3</sub> CN	Acetonitrile
85.02600	[A <sub>1</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
85.05887	[M+H] <sup>+</sup>	C <sub>2</sub> D <sub>6</sub> OS	d6-DMSO
88.03931	[A <sub>1</sub> B <sub>1</sub> +H] <sup>+</sup>	(CH <sub>3</sub> CN) <sub>n</sub> (HCOOH) <sub>m</sub>	Acetonitrile/Formic Acid
96.04198	[A <sub>1</sub> B <sub>1</sub> +Na] <sup>+</sup>	(CH <sub>3</sub> CN) <sub>n</sub> (CH <sub>3</sub> OH) <sub>m</sub>	Acetonitrile/Methanol
99.04165	[A <sub>1</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
100.07569	[M+H] <sup>+</sup>	C <sub>5</sub> H <sub>10</sub> NO	NMP
100.99994	[A <sub>1</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
101.00316	[M+Na] <sup>+</sup>	C <sub>2</sub> H <sub>6</sub> OS	DMSO
101.08084	[A <sub>2</sub> B <sub>2</sub> +H] <sup>+</sup>	[MeOH] <sub>n</sub> [H <sub>2</sub> O] <sub>m</sub>	Methanol/Water
102.05496	[A <sub>1</sub> B <sub>1</sub> +H] <sup>+</sup>	(CH <sub>3</sub> CN) <sub>n</sub> (CH <sub>3</sub> COOH) <sub>m</sub>	Acetonitrile/Acetic Acid
102.12773	[M+H] <sup>+</sup>	C <sub>6</sub> H <sub>15</sub> N	TEA
103.95560	[M+H] <sup>+</sup>	C <sub>2</sub> H <sub>3</sub> N	ACN
104.99229	[M+Na] <sup>+</sup>	C <sub>2</sub> H <sub>3</sub> O <sub>2</sub> Na	Sodium acetate
105.04232	[M <sub>2</sub> +Na] <sup>+</sup>	C <sub>2</sub> H <sub>3</sub> N	ACN
105.95379	[M+ <sup>65</sup> Cu] <sup>+</sup>	C <sub>2</sub> H <sub>3</sub> N	ACN
107.07027	[A <sub>2</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
115.01559	[A <sub>1</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
115.08659	[A <sub>1</sub> B <sub>1</sub> +H] <sup>+</sup>	(CH <sub>3</sub> CN) <sub>n</sub> (C <sub>3</sub> H <sub>7</sub> NO) <sub>m</sub>	Acetonitrile/Dimethylformamide
120.04776	[M+CH <sub>3</sub> CN+H] <sup>+</sup>	C <sub>2</sub> H <sub>6</sub> OS	DMSO
122.08117	[M+H] <sup>+</sup>	C <sub>4</sub> H <sub>11</sub> NO <sub>3</sub>	TRIS
123.06278	[A <sub>2</sub> B <sub>2</sub> +Na] <sup>+</sup>	[MeOH] <sub>n</sub> [H <sub>2</sub> O] <sub>m</sub>	Methanol/Water
123.09167	[M+H] <sup>+</sup>	C <sub>7</sub> H <sub>10</sub> N <sub>2</sub>	DMAP
124.03690	[A <sub>1</sub> B <sub>1</sub> +Na] <sup>+</sup>	(CH <sub>3</sub> CN) <sub>n</sub> (CH <sub>3</sub> COOH) <sub>m</sub>	Acetonitrile/Acetic Acid
129.05222	[A <sub>2</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
130.15903	[M+H] <sup>+</sup>	C <sub>8</sub> H <sub>19</sub> N	DIPEA
132.90490	M <sup>+</sup>	Cs	Cs-133
133.10705	[A <sub>3</sub> B <sub>2</sub> +H] <sup>+</sup>	[MeOH] <sub>n</sub> [H <sub>2</sub> O] <sub>m</sub>	Methanol/Water
135.10157	[A <sub>2</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
137.07431	[M+CH <sub>3</sub> CN+NH <sub>4</sub> ] <sup>+</sup>	C <sub>2</sub> H <sub>6</sub> OS	DMSO
142.02971	[M+CH <sub>3</sub> CN+Na] <sup>+</sup>	C <sub>2</sub> H <sub>6</sub> OS	DMSO
144.17468	[M+H] <sup>+</sup>	C <sub>9</sub> H <sub>21</sub> N	TPA
144.98215	[M <sub>2</sub> + <sup>63</sup> Cu] <sup>+</sup>	CH <sub>3</sub> CN	ACN
145.02615	[A <sub>2</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
146.06887	[M <sub>3</sub> +Na] <sup>+</sup>	CH <sub>3</sub> CN	ACN
146.98034	[M <sub>2</sub> + <sup>65</sup> Cu] <sup>+</sup>	CH <sub>3</sub> CN	ACN

## Appendix

Ion Mass	Ion Type	Formula or Subunit	Species
147.11280	$[A_2B_2+H]^+$	$(CH_3CN)_n(CH_3OH)_m$	Acetonitrile/Methanol
149.02332	$[f+H]^+$	$C_8H_4O_3$	Phthalic Anhydride
150.12773	$[M+H]^+$	$C_{10}H_{15}N$	Phenyldiethylamine
151.09649	$[A_3B+H]^+$	$[C_2H_4O]_nH_2O$	PEG
153.13862	$[M+H]^+$	$C_9H_{16}N_2$	DBU
155.08900	$[A_3B_2+Na]^+$	$[MeOH]_n[H_2O]_m$	Methanol/Water
157.03515	$[M_2+H]^+$	$C_2H_6OS$	DMSO
157.08352	$[A_2B+Na]^+$	$[C_3H_6O]_nH_2O$	PPG
158.96403	$[M+Na]^+$	$C_2F_3O_2Na$	NaTFA
163.03897	$[M-CH_3OH+H]^+$	$C_{10}H_{10}O_4$	Dimethyl phthalate
163.13287	$[M+H]^+$	$C_8H_{18}O_3$	DGBE
169.09475	$[A_2B_2+Na]^+$	$(CH_3CN)_n(CH_3OH)_m$	Acetonitrile/Methanol
169.11046	$[M_2+H]^+$	$C_2D_6OS$	d6-DMSO
171.00527	$[f+Na]^+$	$C_8H_4O_3$	Phthalic anhydride
172.03931	$[M-H_2O+H]^+$	$C_{10}H_7NO_3$	4-HCCA
173.05745	$[A_2B+K]^+$	$[C_3H_6O]_nH_2O$	PPG
173.07843	$[A_3B+Na]^+$	$[C_2H_4O]_nH_2O$	PEG
179.01709	$[M_2+Na]^+$	$C_2H_6OS$	DMSO
181.12231	$[M+H]^+$	$C_{11}H_{16}O_2$	BHA
183.08044	$[M+H]^+$	$C_{13}H_{10}O$	DPK
183.14383	$[A_4B_3+H]^+$	$[MeOH]_n[H_2O]_m$	Methanol/Water
185.11482	$[M+Na]^+$	$C_8H_{18}O_3$	GE
186.22163	$[M+H]^+$	$C_{12}H_{27}N$	TBA
189.05237	$[A_3B+K]^+$	$[C_2H_4O]_nH_2O$	PEG
190.04987	$[M+H]^+$	$C_{10}H_7NO_3$	4-HCCA
193.14344	$[A_3B+H]^+$	$[C_3H_6O]_nH_2O$	PPG
195.06519	$[M+H]^+$	$C_{10}H_{10}O_4$	Dimethyl phthalate
195.12270	$[A_4B+H]^+$	$[C_2H_4O]_nH_2O$	PEG
203.10425	$[M+Na]^+$	$C_{11}H_{16}O_2$	BHA
205.12578	$[A_4B_3+Na]^+$	$[MeOH]_n[H_2O]_m$	Methanol/Water
212.03181	$[M+Na]^+$	$C_{10}H_7NO_3$	4-HCCA
214.08963	$[M+H]^+$	$C_{10}H_{15}NO_2S$	n-BBS
215.12538	$[A_3B+Na]^+$	$[C_3H_6O]_nH_2O$	PPG
217.10465	$[A_4B+Na]^+$	$[C_2H_4O]_nH_2O$	PEG
221.18999	$[M+H]^+$	$C_{15}H_{24}O$	BTH
225.19614	$[M+H]^+$	$C_{13}H_{24}N_2O$	DCU
228.00575	$[M+K]^+$	$C_{10}H_7NO_3$	4-HCCA
231.09932	$[A_3B+K]^+$	$[C_3H_6O]_nH_2O$	PPG
231.11618	$[M+NH_4]^+$	$C_{10}H_{15}NO_2S$	n-BBS
233.07858	$[A_4B+K]^+$	$[C_2H_4O]_nH_2O$	PEG
236.07157	$[M+Na]^+$	$C_{10}H_{15}NO_2S$	n-BBS
239.14892	$[A_5B+H]^+$	$[C_2H_4O]_nH_2O$	PEG
239.22485	$[(M.H_{35}Cl)_2-Cl]^+$	$C_6H_{15}N$	TEA.HCl
241.22190	$[(M.H_{37}Cl)_2-Cl]^+$	$C_6H_{15}N$	TEA.HCl
242.28423	$M^+$	$C_{16}H_{36}N$	TBA
243.11683	$M^+$	$C_{19}H_{15}$	Trityl cation
243.17194	$[M+Na]^+$	$C_{15}H_{24}O$	BTH
251.18530	$[A_4B+H]^+$	$[C_3H_6O]_nH_2O$	PPG
251.20056	$[AB_1+H]^+$	$[C_{14}H_{22}O][C_2H_4O]_n$	Triton
257.03103	$[M_3+Na]^+$	$C_2H_6OS$	DMSO
261.13086	$[A_5B+Na]^+$	$[C_2H_4O]_nH_2O$	PEG
265.21621	$[AB_1+H]^+$	$[C_{15}H_{24}O][C_2H_4O]_n$	Triton
267.17197	$[M+H]^+$	$C_{12}H_{27}O_4P$	TBP
273.12739	$M^+$	$C_{20}H_{17}O$	MMT
273.16725	$[A_4B+Na]^+$	$[C_3H_6O]_nH_2O$	PPG
273.18250	$[AB_1+Na]^+$	$[C_{14}H_{22}O][C_2H_4O]_n$	Triton

## Appendix

Ion Mass	Ion Type	Formula or Subunit	Species
277.10480	[A <sub>5</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
279.09333	[M+H] <sup>+</sup>	C <sub>18</sub> H <sub>15</sub> OP	TPO
279.15909	[M+H] <sup>+</sup>	C <sub>16</sub> H <sub>22</sub> O <sub>4</sub>	Dibutylphthalate
279.22945	[AB <sub>1</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
282.27914	[M+H] <sup>+</sup>	C <sub>18</sub> H <sub>35</sub> NO	Oleamide
283.17513	[A <sub>6</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
284.29479	[M+H] <sup>+</sup>	C <sub>18</sub> H <sub>37</sub> NO	Stearamide
287.19815	[AB <sub>1</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
288.25332	[M+H] <sup>+</sup>	C <sub>16</sub> H <sub>33</sub> NO <sub>3</sub>	n,n-DDA
289.14118	[A <sub>4</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
293.24510	[AB <sub>1</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
295.22677	[AB <sub>2</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
301.14103	[M+Na] <sup>+</sup>	C <sub>16</sub> H <sub>22</sub> O <sub>4</sub>	Dibutylphthalate
304.26108	[M+Na] <sup>+</sup>	C <sub>18</sub> H <sub>35</sub> NO	Oleamide
305.15708	[A <sub>6</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
306.27673	[M+Na] <sup>+</sup>	C <sub>18</sub> H <sub>37</sub> NO	Stearamide
309.22717	[A <sub>5</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
309.24242	[AB <sub>2</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
315.25299	[M+H] <sup>+</sup>	C <sub>18</sub> H <sub>34</sub> O <sub>4</sub>	DBS
317.11497	[M+K] <sup>+</sup>	C <sub>16</sub> H <sub>22</sub> O <sub>4</sub>	Dibutylphthalate
317.20872	[AB <sub>2</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
321.13101	[A <sub>6</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
323.25567	[AB <sub>2</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
325.25847	[M <sub>2</sub> +H] <sup>+</sup>	C <sub>8</sub> H <sub>18</sub> O <sub>3</sub>	DGBE
327.07807	[M+H] <sup>+</sup>	C <sub>18</sub> H <sub>15</sub> O <sub>4</sub> P	TPP
327.20135	[A <sub>7</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
331.20911	[A <sub>5</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
331.22437	[AB <sub>2</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
337.11841	[M+H] <sup>+</sup> ; ( <sub>120</sub> Sn) <sup>+</sup>	C <sub>13</sub> H <sub>28</sub> O <sub>2</sub> Sn	Tributyl tin formate
337.27132	[AB <sub>2</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
338.34174	[M+H] <sup>+</sup>	C <sub>22</sub> H <sub>43</sub> NO	Erucamide
339.25299	[AB <sub>3</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
347.18305	[A <sub>5</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
349.18329	[A <sub>7</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
353.26864	[AB <sub>3</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
355.06994	[M+H-CH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>5</sub>	Polysiloxane
355.36829	[M-Cl] <sup>+</sup>	C <sub>22</sub> H <sub>47</sub> N <sub>2</sub> OCl	
360.32368	[M+Na] <sup>+</sup>	C <sub>22</sub> H <sub>43</sub> NO	Erucamide
361.23493	[AB <sub>3</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
365.15723	[A <sub>7</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
367.26903	[A <sub>6</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
367.28188	[AB <sub>3</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
368.42508	[M-Cl] <sup>+</sup>	C <sub>25</sub> H <sub>54</sub> NCl	BTAC-228
371.10124	[M+H] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>5</sub>	Polysiloxane
371.22756	[A <sub>8</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
371.31559	[M+H] <sup>+</sup>	C <sub>22</sub> H <sub>42</sub> O <sub>4</sub>	DEHA
371.31559	[M+H] <sup>+</sup>	C <sub>22</sub> H <sub>42</sub> O <sub>4</sub>	DOA
375.25058	[AB <sub>3</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
379.09246	[M <sub>2</sub> +H] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
381.29753	[AB <sub>3</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
383.27920	[AB <sub>4</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
388.12779	[M+NH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>5</sub>	Polysiloxane
389.25098	[A <sub>6</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
391.28429	[M+H] <sup>+</sup>	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	Diisooctyl phthalate
393.20951	[A <sub>8</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
397.29485	[AB <sub>4</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton

## Appendix

Ion Mass	Ion Type	Formula or Subunit	Species
405.22491	[A <sub>6</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
405.26115	[AB <sub>4</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
409.18344	[A <sub>8</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
411.30810	[AB <sub>4</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
413.26623	[M+Na] <sup>+</sup>	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	Diisooctyl phthalate
415.25378	[A <sub>9</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
419.27680	[AB <sub>4</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
425.31090	[A <sub>7</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
425.32375	[AB <sub>4</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
427.30542	[AB <sub>5</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
429.08873	[M+H - CH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>6</sub>	Polysiloxane
429.24017	[M+K] <sup>+</sup>	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	Diisooctyl phthalate
437.23572	[A <sub>9</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
441.01479	[M <sub>3</sub> + <sub>63</sub> Cu(I)] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
441.32107	[AB <sub>5</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
443.01298	[M <sub>3</sub> + <sub>65</sub> Cu(I)] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
445.12003	[M+H] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>6</sub>	Polysiloxane
447.29284	[M+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
449.28736	[AB <sub>5</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
449.38500	[M <sub>2</sub> +H] <sup>+</sup>	C <sub>13</sub> H <sub>24</sub> N <sub>2</sub> O	DCU
453.20966	[A <sub>9</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
453.34353	[M+H] <sup>+</sup>	C <sub>24</sub> H <sub>44</sub> N <sub>4</sub> O <sub>4</sub>	nylon
454.29278	[M+CH <sub>3</sub> CN+Na] <sup>+</sup>	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	Diisooctyl phthalate
455.33431	[AB <sub>5</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
459.27999	[A <sub>10</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
462.14658	[M+NH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>6</sub>	Polysiloxane
463.26678	[A <sub>7</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
463.30301	[AB <sub>5</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
469.34996	[AB <sub>5</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
471.33163	[AB <sub>6</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
472.28781	[M+H] <sup>+</sup>	SLPR	Peptide
481.26194	[A <sub>10</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
483.35276	[A <sub>8</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
485.34728	[AB <sub>6</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
493.31358	[AB <sub>6</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
494.56593	[M - Cl] <sup>+</sup>	C <sub>34</sub> H <sub>72</sub> NCl	DPDMA
497.23587	[A <sub>10</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
499.36053	[AB <sub>6</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
503.10752	[M+H - CH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>7</sub>	Polysiloxane
503.30621	[A <sub>11</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
505.33471	[A <sub>8</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
507.32923	[AB <sub>6</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
513.37618	[AB <sub>6</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
515.33001	[M+H] <sup>+</sup>	IQVR	Peptide
515.35785	[AB <sub>7</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
515.41286	[M+H] <sup>+</sup>	C <sub>30</sub> H <sub>58</sub> O <sub>4</sub> S	DDTDP
519.13882	[M+H] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>7</sub>	Polysiloxane
521.30864	[A <sub>8</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
522.59723	[M - Cl] <sup>+</sup>	C <sub>36</sub> H <sub>76</sub> NCl	SPDMA
525.28815	[A <sub>11</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
529.37350	[AB <sub>7</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
531.40777	[M+H] <sup>+</sup>	C <sub>30</sub> H <sub>58</sub> O <sub>5</sub> S	DDTDP
531.47717	[M+H] <sup>+</sup>	C <sub>35</sub> H <sub>62</sub> O <sub>3</sub>	Irganox
536.16537	[M+NH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>7</sub>	Polysiloxane
537.33979	[AB <sub>7</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
537.87901	[M <sub>6</sub> - <sub>6</sub> H+ <sub>3</sub> Fe+O] <sup>+</sup>	C <sub>2</sub> H <sub>4</sub> O <sub>2</sub>	Acetic acid-Fe-O- complex

## Appendix

Ion Mass	Ion Type	Formula or Subunit	Species
541.26209	[A <sub>11</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
541.39463	[A <sub>9</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
543.38674	[AB <sub>7</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
547.33242	[A <sub>12</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
547.40269	[M+H] <sup>+</sup>	C <sub>30</sub> H <sub>58</sub> O <sub>6</sub> S	DDTDP
550.62853	[M - Cl] <sup>+</sup>	C <sub>38</sub> H <sub>80</sub> NCI	DSDMA
551.35544	[AB <sub>7</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
553.38972	[M+Na] <sup>+</sup>	C <sub>30</sub> H <sub>58</sub> O <sub>5</sub> S	DDTDP
553.45912	[M+Na] <sup>+</sup>	C <sub>35</sub> H <sub>62</sub> O <sub>3</sub>	Irganox
557.40239	[AB <sub>7</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
559.38406	[AB <sub>8</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
563.37657	[A <sub>9</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
568.13506	[M <sub>3</sub> +H] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
569.31437	[A <sub>12</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
571.35622	[M+H] <sup>+</sup>	VSLPR	Peptide
573.39971	[AB <sub>8</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
577.12631	[M+H - CH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>8</sub>	Polysiloxane
579.35051	[A <sub>9</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
581.36601	[AB <sub>8</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
585.28830	[A <sub>12</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
587.41296	[AB <sub>8</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
591.35864	[A <sub>13</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
593.15761	[M+H] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>8</sub>	Polysiloxane
595.38166	[AB <sub>8</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
599.43649	[A <sub>10</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
601.42861	[AB <sub>8</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
603.41028	[AB <sub>9</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
606.09149	[M <sub>3</sub> +K] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
610.18416	[M+NH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>8</sub>	Polysiloxane
613.34058	[A <sub>13</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
615.40375	[M+H] <sup>+</sup>	C <sub>32</sub> H <sub>58</sub> N <sub>2</sub> O <sub>7</sub> S	CHAPS
617.42593	[AB <sub>9</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
621.41844	[A <sub>10</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
621.97291	[M <sub>6</sub> - <sub>6</sub> H + <sub>3</sub> Fe + O] <sup>+</sup>	C <sub>3</sub> H <sub>6</sub> O <sub>2</sub>	Propionic acid Fe-O complex
625.39222	[AB <sub>9</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
629.31452	[A <sub>13</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
631.43917	[AB <sub>9</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
633.32023	[M+H] <sup>+</sup>	QTIASN	Peptide
635.38485	[A <sub>14</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
637.39237	[A <sub>10</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
639.40787	[AB <sub>9</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
645.45482	[AB <sub>9</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
647.43649	[AB <sub>10</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
651.14510	[M+H - CH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>9</sub>	Polysiloxane
657.36680	[A <sub>14</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
657.47836	[A <sub>11</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
659.38350	[M+H] <sup>+</sup>	SGIQVR	Peptide
661.45214	[AB <sub>10</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
667.17640	[M+H] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>9</sub>	Polysiloxane
669.41844	[AB <sub>10</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
672.40390	[M+H] <sup>+</sup>	TVSLPR	Peptide
672.40390	[M+H] <sup>+</sup>	TVSLPR	Peptide
673.34073	[A <sub>14</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
675.46539	[AB <sub>10</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
679.41107	[A <sub>15</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
679.46030	[A <sub>11</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG

## Appendix

Ion Mass	Ion Type	Formula or Subunit	Species
679.51166	[M+H] <sup>+</sup>	C <sub>36</sub> H <sub>66</sub> N <sub>6</sub> O <sub>6</sub>	nylon
683.43409	[AB <sub>10</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
684.20295	[M+NH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>9</sub>	Polysiloxane
689.48104	[AB <sub>10</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
691.46271	[AB <sub>11</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
695.43424	[A <sub>11</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
701.39301	[A <sub>15</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
704.38250	[M+H] <sup>+</sup>	LDSELK	Peptide
705.47836	[AB <sub>11</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
713.44465	[AB <sub>11</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
715.52022	[A <sub>12</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
717.36695	[A <sub>15</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
719.49160	[AB <sub>11</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
723.43728	[A <sub>16</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
725.16390	[M+H-CH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>10</sub>	Polysiloxane
727.46030	[AB <sub>11</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
732.46544	[M+H] <sup>+</sup>	GLVLI AF	Peptide
733.50725	[AB <sub>11</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
735.48892	[AB <sub>12</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
737.50217	[A <sub>12</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
741.19520	[M+H] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>10</sub>	Polysiloxane
742.44979	[M+H] <sup>+</sup>	GPFPI LV	Peptide
743.44101	[M+H] <sup>+</sup>	ATVSLPR	Peptide
745.41923	[A <sub>16</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
749.50457	[AB <sub>12</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
753.47610	[A <sub>12</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
757.47087	[AB <sub>12</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
758.22175	[M+NH <sub>4</sub> ] <sup>+</sup>	[C <sub>2</sub> H <sub>6</sub> SiO] <sub>10</sub>	Polysiloxane
758.41553	[M+H] <sup>+</sup>	PATLNSR	Peptide
761.39316	[A <sub>16</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
763.51782	[AB <sub>12</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
767.46350	[A <sub>17</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
771.48652	[AB <sub>12</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
773.56209	[A <sub>13</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
777.53347	[AB <sub>12</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
779.51514	[AB <sub>13</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
789.44544	[A <sub>17</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
793.53079	[AB <sub>13</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
795.54403	[A <sub>13</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
798.58785	[M <sub>2</sub> +NH <sub>4</sub> ] <sup>+</sup>	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	Diisooctyl phthalate
801.49708	[AB <sub>13</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
802.43051	[M+H] <sup>+</sup>	LSSPATLN	Peptide
803.54324	[M <sub>2</sub> +Na] <sup>+</sup>	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	Diisooctyl phthalate
804.40978	[M+H] <sup>+</sup>	SEIDNVK	Peptide
805.41626	[M+H] <sup>+</sup>	SAASLNSR	Peptide
805.41938	[A <sub>17</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
807.39954	[M+H] <sup>+</sup>	LAADDFR	Peptide
807.54403	[AB <sub>13</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
809.44035	[M+H] <sup>+</sup>	LASYLDK	Peptide
809.48691	[AB <sub>10</sub> +Na] <sup>+</sup>	[C <sub>18</sub> H <sub>34</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
811.48971	[A <sub>18</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
811.51797	[A <sub>13</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
815.51273	[AB <sub>13</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
819.51718	[M <sub>2</sub> +K] <sup>+</sup>	C <sub>24</sub> H <sub>38</sub> O <sub>4</sub>	Diisooctyl phthalate
821.55968	[AB <sub>13</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
823.54135	[AB <sub>14</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton

## Appendix

Ion Mass	Ion Type	Formula or Subunit	Species
824.49887	[M+H] <sup>+</sup>	PGVVSLPR	Peptide
827.42978	[M+H] <sup>+</sup>	FASFIDK	Peptide
827.46214	[M+H] <sup>+</sup>	PEIQNVK	Peptide
831.60395	[A <sub>14</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
832.48870	[M+H] <sup>+</sup>	SISISVAR	Peptide
833.47166	[A <sub>18</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
837.55700	[AB <sub>14</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
839.09742	[M <sub>4</sub> - <sub>2</sub> H+K+ <sub>2</sub> Na] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
842.50943	[M+H] <sup>+</sup>	VATVSLPR	Peptide
845.10543	[M <sub>4</sub> - <sub>3</sub> H+ <sub>4</sub> Na] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
845.52330	[AB <sub>14</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
848.49886	[M+H] <sup>+</sup>	AFIDKVR	Peptide
849.44559	[A <sub>18</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
851.57025	[AB <sub>14</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
853.51313	[AB <sub>11</sub> +Na] <sup>+</sup>	[C <sub>18</sub> H <sub>34</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
853.58590	[A <sub>14</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
855.07136	[M <sub>4</sub> - <sub>2</sub> H+Na+ <sub>2</sub> K] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
855.51593	[A <sub>19</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
859.53895	[AB <sub>14</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
861.07937	[M <sub>4</sub> - <sub>3</sub> H+ <sub>3</sub> Na+K] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
865.54951	[AB <sub>10</sub> +Na] <sup>+</sup>	[C <sub>22</sub> H <sub>42</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
865.58590	[AB <sub>14</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
867.08737	[M <sub>4</sub> - <sub>4</sub> H+ <sub>5</sub> Na] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
867.56757	[AB <sub>15</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
869.55983	[A <sub>14</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
870.54073	[M+H] <sup>+</sup>	VATVSLPRN-term-methylated	Peptide
871.04530	[M <sub>4</sub> - <sub>2</sub> H+ <sub>3</sub> K] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
871.49959	[M+H] <sup>+</sup>	QATVSLPR	Peptide
874.49926	[M+H] <sup>+</sup>	SLVNLGGSK	Peptide
877.49787	[A <sub>19</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
881.47271	[M+H] <sup>+</sup>	SLYGLGGSK	Peptide
881.58322	[AB <sub>15</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
883.51485	[M+H] <sup>+</sup>	RVYVHPI	Peptide
889.54951	[AB <sub>15</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
889.64582	[A <sub>15</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
891.56516	[AB <sub>10</sub> +Na] <sup>+</sup>	[C <sub>24</sub> H <sub>44</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
893.47181	[A <sub>19</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
893.58081	[AB <sub>10</sub> +Na] <sup>+</sup>	[C <sub>24</sub> H <sub>46</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
895.59646	[AB <sub>15</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
897.53934	[AB <sub>12</sub> +Na] <sup>+</sup>	[C <sub>18</sub> H <sub>34</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
899.53089	[M+H] <sup>+</sup>	VQTVSLPR	Peptide
899.54214	[A <sub>20</sub> B+H] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
903.56516	[AB <sub>15</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
905.67979	[M+H] <sup>+</sup>	C <sub>48</sub> H <sub>88</sub> N <sub>8</sub> O <sub>8</sub>	nylon
906.50434	[M+H] <sup>+</sup>	NKPGVYTK	Peptide
906.50434	[M+H] <sup>+</sup>	NKPGVYTK	Peptide
909.57573	[AB <sub>11</sub> +Na] <sup>+</sup>	[C <sub>22</sub> H <sub>42</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
909.61211	[AB <sub>15</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
911.59378	[AB <sub>16</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
911.62776	[A <sub>15</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
917.49920	[M+H] <sup>+</sup>	RVYVHPF	Peptide
921.52409	[A <sub>20</sub> B+Na] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
925.60943	[AB <sub>16</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
927.60170	[A <sub>15</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
931.51485	[M+H] <sup>+</sup>	RVYIHPF	Peptide
933.57573	[AB <sub>16</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton



## Appendix

Ion Mass	Ion Type	Formula or Subunit	Species
935.59138	[AB <sub>11</sub> +Na] <sup>+</sup>	[C <sub>24</sub> H <sub>44</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
937.49802	[A <sub>20</sub> B+K] <sup>+</sup>	[C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub> H <sub>2</sub> O	PEG
937.60703	[AB <sub>11</sub> +Na] <sup>+</sup>	[C <sub>24</sub> H <sub>46</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
939.62268	[AB <sub>16</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
941.56556	[AB <sub>13</sub> +Na] <sup>+</sup>	[C <sub>18</sub> H <sub>34</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
947.59138	[AB <sub>16</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
947.68768	[A <sub>16</sub> B+H] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
950.47305	[M+H] <sup>+</sup>	YVNWIQQ	Peptide
953.60194	[AB <sub>12</sub> +Na] <sup>+</sup>	[C <sub>22</sub> H <sub>42</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
953.63833	[AB <sub>16</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
955.62000	[AB <sub>17</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
969.63565	[AB <sub>17</sub> +H] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
969.66963	[A <sub>16</sub> B+Na] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
973.53129	[M+H] <sup>+</sup>	IEISELNR	Peptide
977.60194	[AB <sub>17</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
979.50949	[M+H] <sup>+</sup>	GTSYDPVLK	Peptide
979.61759	[AB <sub>12</sub> +Na] <sup>+</sup>	[C <sub>24</sub> H <sub>44</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
981.63324	[AB <sub>12</sub> +Na] <sup>+</sup>	[C <sub>24</sub> H <sub>46</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
983.64889	[AB <sub>17</sub> +Na] <sup>+</sup>	[C <sub>14</sub> H <sub>28</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
985.59177	[AB <sub>14</sub> +Na] <sup>+</sup>	[C <sub>18</sub> H <sub>34</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
985.64356	[A <sub>16</sub> B+K] <sup>+</sup>	[C <sub>3</sub> H <sub>6</sub> O] <sub>n</sub> H <sub>2</sub> O	PPG
991.61759	[AB <sub>17</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>24</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton
994.15551	[M <sub>5</sub> - H <sub>2</sub> O - <sub>2</sub> H + <sub>3</sub> Na] <sup>+</sup>	C <sub>10</sub> H <sub>7</sub> NO <sub>3</sub>	4-HCCA
995.51966	[M+H] <sup>+</sup>	IKEWYEK	Peptide
997.62816	[AB <sub>13</sub> +Na] <sup>+</sup>	[C <sub>22</sub> H <sub>42</sub> O <sub>6</sub> ][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Tween
997.66454	[AB <sub>17</sub> +Na] <sup>+</sup>	[C <sub>15</sub> H <sub>30</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton, reduced
999.64621	[AB <sub>18</sub> +H] <sup>+</sup>	[C <sub>14</sub> H <sub>22</sub> O][C <sub>2</sub> H <sub>4</sub> O] <sub>n</sub>	Triton

## Appendix

# LC-MS Adducts, Losses, and Replacements

List of potential gains, losses, and replacements in positive ion mode LC-ESI-MS. The list is adapted from Keller *et al.*<sup>[110]</sup>. The exact mass difference, ion type, reaction, and potential origin of the compound is provided.

Mass Difference	Ion type	Reaction	Origin
14.01565	$[A: [C_3H_6ON] \leftrightarrow [C_2H_4ON]]^+$	$C_3H_6ON \rightleftharpoons C_2H_4ON$	acrylamide/iodoacetamide
43.94948442	$[A: ^{79}Br \leftrightarrow Cl]^+$	$^{79}Br \rightleftharpoons Cl$	halogen exchange
52.9152631	$[A: ^{79}Br \leftrightarrow CN]^+$	$^{79}Br \rightleftharpoons CN$	halogen exchange
77.9105121	$[A: ^{79}Br \leftrightarrow H]^+$	$^{79}Br \rightleftharpoons H$	halogen exchange
61.9155971	$[A: ^{79}Br \leftrightarrow OH]^+$	$^{79}Br \rightleftharpoons OH$	halogen exchange
45.94743792	$[A: ^{81}Br \leftrightarrow Cl]^+$	$^{81}Br \rightleftharpoons Cl$	halogen exchange
54.9132166	$[A: ^{81}Br \leftrightarrow CN]^+$	$^{81}Br \rightleftharpoons CN$	halogen exchange
79.9084656	$[A: ^{81}Br \leftrightarrow H]^+$	$^{81}Br \rightleftharpoons H$	halogen exchange
63.9135506	$[A: ^{81}Br \leftrightarrow OH]^+$	$^{81}Br \rightleftharpoons OH$	halogen exchange
37.989162	$[A: 2CN \leftrightarrow 2COOH]^+$	$2CN \rightleftharpoons 2COOH$	nitrile compounds
8.96577868	$[A: Cl \leftrightarrow CN]^+$	$Cl \rightleftharpoons CN$	halogen exchange
33.96102768	$[A: Cl \leftrightarrow H]^+$	$Cl \rightleftharpoons H$	halogen exchange
17.96611268	$[A: Cl \leftrightarrow OH]^+$	$Cl \rightleftharpoons OH$	halogen exchange
18.994581	$[A: CN \leftrightarrow COO]^+$	$CN \rightleftharpoons COO$	nitrile compounds
24.995249	$[A: CN \leftrightarrow H]^+$	$CN \rightleftharpoons H$	nitrile compounds
7.00467078	$[A: F \leftrightarrow CN]^+$	$F \rightleftharpoons CN$	halogen exchange
17.99057822	$[A: F \leftrightarrow H]^+$	$F \rightleftharpoons H$	halogen exchange
1.99566322	$[A: F \leftrightarrow OH]^+$	$F \rightleftharpoons OH$	halogen exchange
91.93562072	$[A: I \leftrightarrow Cl]^+$	$I \rightleftharpoons Cl$	halogen exchange
100.9013994	$[A: I \leftrightarrow CN]^+$	$I \rightleftharpoons CN$	halogen exchange
125.8966484	$[A: I \leftrightarrow H]^+$	$I \rightleftharpoons H$	halogen exchange
109.9017334	$[A: I \leftrightarrow OH]^+$	$I \rightleftharpoons OH$	halogen exchange
37.955881	$[A: K^+ \leftrightarrow H^+]^+$	$K^+ \rightleftharpoons H^+$	salt adduct
20.929332	$[A: K^+ \leftrightarrow NH_4^+]^+$	$K^+ \rightleftharpoons NH_4^+$	salt adduct
21.981944	$[A: Na^+ \leftrightarrow H^+]^+$	$Na^+ \rightleftharpoons H^+$	salt adduct
15.973937	$[A: Na^+ \leftrightarrow K^+]^+$	$Na^+ \rightleftharpoons K^+$	salt adduct
4.955395	$[A: Na^+ \leftrightarrow NH_4^+]^+$	$Na^+ \rightleftharpoons NH_4^+$	salt adduct
17.026549	$[A: NH_4^+ \leftrightarrow H^+]^+$	$NH_4^+ \rightleftharpoons H^+$	salt adduct
29.97418	$[A: NO_2 \leftrightarrow NH_2]^+$	$NO_2 \rightleftharpoons NH_2$	nitro compounds
13.979265	$[A: O \leftrightarrow 2H]^+$	$O \rightleftharpoons 2H$	Oxidation
0.984016	$[A: OH \leftrightarrow NH_2]^+$	$OH \rightleftharpoons NH_2$	de-amidation
15.977156	$[A: S \leftrightarrow O]^+$	$S \rightleftharpoons O$	sulfur compounds
28.006148	$[A - 2N]^+$		nitrogen loss
63.998286	$[A - CH_3SOH]^+$		oxidized methionines
33.021464	$[A - NH_2OH]^+$		hydroxamic acids
29.997989	$[A - NO]^+$		nitroso compounds
2.01565	$[A \pm 2H]^+$		double bond formation
31.98983	$[A \pm 2O]^+$		oxygen loss
305.068158	$[A \pm C_{10}O_6N_3SH_{15}]^+$		glutathione+o-water
307.083808	$[A \pm C_{10}O_6N_3SH_{17}]^+$		glutathione
289.073243	$[A \pm C_{10}O_5N_3SH_{15}]^+$		glutathione-water
291.095419	$[A \pm C_{11}O_8NH_{17}]^+$		sialic acid
309.105984	$[A \pm C_{11}O_9NH_{19}]^+$		sialic acid

## Appendix

Mass Difference	Ion type	Reaction	Origin
324.10565	$[A \pm C_{12}O_{10}H_{20}]^+$		sucrose-water
342.116215	$[A \pm C_{12}O_{11}H_{22}]^+$		sucrose
28.0313	$[A \pm C_2H_4]^+$		natural alkane chains
42.04695	$[A \pm C_3H_6]^+$		propylation
56.0626	$[A \pm C_4H_8]^+$		butylation
146.05791	$[A \pm C_6O_4H_{10}]^+$		deoxy-hexose-water
162.052825	$[A \pm C_6O_5H_{10}]^+$		hexose-water
164.068475	$[A \pm C_6O_5H_{12}]^+$		deoxy-hexose-water
180.06339	$[A \pm C_6O_6H_{12}]^+$		hexose
176.03209	$[A \pm C_6O_6H_8]^+$		glucuronic acid
194.042655	$[A \pm C_6O_7H_{10}]^+$		glucuronic acid
203.079374	$[A \pm C_8O_5NH_{13}]^+$		n-acetylhexoseamine
221.089939	$[A \pm C_8O_6NH_{15}]^+$		n-acetylhexoseamine
14.01565	$[A \pm CH_2]^+$		methylation
27.994915	$[A \pm CO]^+$		carbon monoxide
58.00548	$[A \pm CO_2CH_2]^+$		ester
43.98983	$[A \pm CO_2]^+$		decarboxylation
42.010565	$[A \pm COCH_2]^+$		acetyl loss/gain
43.005814	$[A \pm CONH]^+$		acyl amide loss/gain
33.987721	$[A \pm H_2S]^+$		sulfur compounds
97.976897	$[A \pm H_3PO_4]^+$		phosphorous compounds
18.010565	$[A \pm H_2O]^+$		water addition/loss
97.967381	$[A \pm H_2SO_4]^+$		sulfur compounds
27.010899	$[A \pm HCN]^+$		nitrile compounds
17.026549	$[A \pm NH_3]^+$		ammonium adduct
15.994915	$[A \pm O]^+$		oxidation/reduction
31.972071	$[A \pm S]^+$		sulfur compounds
47.966986	$[A \pm SO]^+$		sulfur compounds
63.961901	$[A \pm SO_2]^+$		sulfur compounds
79.956816	$[A \pm SO_3]^+$		sulfur compounds
40.0313	$[A + (C_3H_6O - H_2O)]^+$		acetone condensation
58.041865	$[A + C_3H_6O]^+$		acetone condensation



---

# REFERENCES

---

- [1] Ewa Urbanczyk-Wochniak, Alexander Luedemann, Joachim Kopka, Joachim Selbig, Ute Roessner-Tunali, Lothar Willmitzer, and Alisdair R Fernie. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO reports*, 4(10):989–93, October 2003.
- [2] C H Waddington. The epigenotype. *Endeavour*, 1:18–20, 1942.
- [3] Haleem J Issaq, Que N Van, Timothy J Waybright, Gary M Muschik, and Timothy D Veenstra. Analytical and statistical approaches to metabolomics research. *Journal of separation science*, 32(13):2183–99, July 2009.
- [4] Tobias Kind, Kwang-Hyeon Liu, Do Yup Lee, Brian DeFelice, John K Meissen, and Oliver Fiehn. Lipid-Blast in silico tandem mass spectrometry database for lipid identification. *Nature methods*, 10(8):755–8, August 2013.
- [5] Manish Sud, Eoin Fahy, and Shankar Subramaniam. Template-based combinatorial enumeration of virtual compound libraries for lipids. *Journal of Cheminformatics*, 4(1):23, 2012.
- [6] Scott Boyer, Catrin Hasselgren Arnby, Lars Carlsson, James Smith, Viktor Stein, and Robert C Glen. Reaction site mapping of xenobiotic biotransformations. *Journal of chemical information and modeling*, 47(2):583–90, 2007.
- [7] Bryson D Bennett, Elizabeth H Kimball, Melissa Gao, Robin Osterhout, Stephen J Van Dien, and Joshua D Rabinowitz. Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nature chemical biology*, 5(8):593–9, August 2009.
- [8] Neepa Y Choksi, Gloria D Jahnke, Cathy St Hilaire, and Michael Shelby. Role of thyroid hormones in human and laboratory animal reproductive health. *Birth defects research. Part B, Developmental and reproductive toxicology*, 68(6):479–91, December 2003.
- [9] E Grouzmann, M Fathi, M Gillet, a de Torrenté, C Cavadas, H Brunner, and T Buclin. Disappearance rate of catecholamines, total metanephrines, and neuropeptide Y from the plasma of patients after resection of pheochromocytoma. *Clinical chemistry*, 47(6):1075–82, June 2001.
- [10] James R Lupski. Genetics. Genome mosaicism—one human, multiple genomes. *Science (New York, N.Y.)*, 341(6144):358–9, July 2013.
- [11] Xiaojun Feng, Xin Liu, Qingming Luo, and Bi-Feng Liu. Mass spectrometry in systems biology: an overview. *Mass spectrometry reviews*, 27(6):635–60, 2008.
- [12] Katja Dettmer, P.A. Aronov, and B.D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51, 2007.
- [13] Vladimir Shulaev. Metabolomics technology and bioinformatics. *Briefings in bioinformatics*, 7(2):128–39, June 2006.
- [14] Oliver Fiehn. Metabolomics—the link between genotypes and phenotypes. *Plant molecular biology*, 48(1-2):155–71, January 2002.
- [15] David I Ellis, Warwick B Dunn, Julian L Griffin, J William Allwood, and Royston Goodacre. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*, 8(9):1243–66, September 2007.

## REFERENCES

---

- [16] David S Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, Souhaila Bouatra, Igor Sinelnikov, David Arndt, Jianguo Xia, Philip Liu, Faizath Yallou, Trent Bjorndahl, Rolando Perez-Pineiro, Roman Eisner, Felicity Allen, Vanessa Neveu, Russ Greiner, and Augustin Scalbert. HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic acids research*, 41:801–807, 2012.
- [17] Tobias Kind, Martin Scholz, and Oliver Fiehn. How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLoS one*, 4(5):e5440, January 2009.
- [18] Gary J Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews. Molecular cell biology*, 13(4):263–9, April 2012.
- [19] Royston Goodacre. An overflow of . . . what else but metabolism! *Metabolomics*, 6(1):1–2, February 2010.
- [20] Oliver Fiehn, Don Robertson, Jules Griffin, Mariet Werf, Basil Nikolau, Norman Morrison, Lloyd W. Sumner, Roy Goodacre, Nigel W. Hardy, Chris Taylor, Jennifer Fostel, Bruce Kristal, Rima Kaddurah-Daouk, Pedro Mendes, Ben Ommen, John C. Lindon, and Susanna-Assunta Sansone. The metabolomics standards initiative (MSI). *Metabolomics*, 3(3):175–178, August 2007.
- [21] Jocelyn Kaiser. Proteomics. Public-private group maps out initiatives. *Science (New York, N.Y.)*, 296(5569):827, May 2002.
- [22] Souhaila Bouatra, Farid Aziat, Rupasri Mandal, An Chi Guo, Michael R. Wilson, Craig Knox, Trent C. Bjorndahl, Ramanarayan Krishnamurthy, Fozia Saleem, Philip Liu, Zerihun T. Dame, Jenna Poelzer, Jessica Huynh, Faizath S. Yallou, Nick Psychogios, Edison Dong, Ralf Bogumil, Cornelia Roehring, and David S. Wishart. The Human Urine Metabolome. *PLoS ONE*, 8(9):e73076, September 2013.
- [23] Kenneth Haug, Reza M Salek, Pablo Conesa, Janna Hastings, Paula de Matos, Mark Rijnbeek, Tejasvi Mahendrakar, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, Eamonn Maguire, Alejandra González-Beltrán, Susanna-Assunta Sansone, Julian L Griffin, and Christoph Steinbeck. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, 41(Database issue):D781–6, January 2013.
- [24] Manhoi Hur, Alexis Ann Campbell, Marcia Almeida-de Macedo, Ling Li, Nick Ransom, Adarsh Jose, Matt Crispin, Basil J Nikolau, and Eve Syrkin Wurtele. A global approach to analysis and interpretation of metabolic data for plant natural product discovery. *Natural product reports*, 30:565–83, 2013.
- [25] Nikolas Kessler, Heiko Neuweger, Anja Bonte, Georg Langenkämper, Karsten Niehaus, Tim W Nattkemper, and Alexander Goesmann. MeltDB 2.0—advances of the metabolomics software system. *Bioinformatics (Oxford, England)*, 29(19):2452–2459, August 2013.
- [26] Lionel Blanchet, Agnieszka Smolinska, Amos Attali, Marcel P Stoop, Kirsten Am Ampt, Hans van Aken, Ernst Suidgeest, Tinka Tuinstra, Sybren S Wijmenga, Theo Luider, and Lutgarde Mc Buydens. Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. *BMC bioinformatics*, 12(1):254, June 2011.
- [27] J E MacNair, K C Lewis, and J W Jorgenson. Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. *Analytical chemistry*, 69(6):983–9, March 1997.
- [28] Kamlesh D Patel, Anton D Jerkovich, Jason C Link, and James W Jorgenson. In-depth characterization of slurry packed capillary columns with 1.0-microm nonporous particles using reversed-phase isocratic ultrahigh-pressure liquid chromatography. *Analytical chemistry*, 76(19):5777–86, October 2004.
- [29] John V Seeley and Stacy K Seeley. Multidimensional gas chromatography: fundamental advances and new applications. *Analytical chemistry*, 85(2):557–78, January 2013.
- [30] Agnieszka Smolinska, Lionel Blanchet, Lutgarde M C Buydens, and Sybren S Wijmenga. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica chimica acta*, 750:82–97, October 2012.
- [31] Mohamed a. Farag, Andrea Porzel, Jürgen Schmidt, and Ludger a. Wessjohann. Metabolite profiling and fingerprinting of commercial cultivars of *Humulus lupulus* L. (hop): a comparison of MS and NMR methods in metabolomics. *Metabolomics*, 8(3):492–507, August 2011.
- [32] K. a. Aliferis and S. Jabaji. <sup>1</sup>H NMR and GC-MS metabolic fingerprinting of developmental stages of *Rhizoctonia solani* sclerotia. *Metabolomics*, 6(1):96–108, September 2009.
- [33] Arjen Lommen, Arjen Gerssen, J. Efraim Oosterink, Harrie J. Kools, Ainhoa Ruiz-Aracama, Ruud J. B. Peters, and Hans G. J. Mol. Ultra-fast searching assists in evaluating sub-ppm mass accuracy enhancement in U-HPLC/Orbitrap MS data. *Metabolomics : Official journal of the Metabolomic Society*, 7(1):15–24, March 2011.

## REFERENCES

---

- [34] Jurre J Kamphorst, Jing Fan, Wenyun Lu, Eileen White, and Joshua D Rabinowitz. Liquid chromatography-high resolution mass spectrometry analysis of fatty acid metabolism. *Analytical chemistry*, 83(23):9114–22, December 2011.
- [35] Haitao Lv. Mass spectrometry-based metabolomics towards understanding of gene functions with a diversity of biological contexts. *Mass spectrometry reviews*, 32(2):118–28, 2012.
- [36] Sofia Moco, Jacques Vervoort, Raoul J. Bino, Ric C.H. De Vos, and Raoul Bino. Metabolomics technologies and metabolite identification. *TrAC Trends in Analytical Chemistry*, 26(9):855–866, October 2007.
- [37] Bin Zhou, Jun Feng Xiao, Leepika Tuli, and Habtom W Ressom. LC-MS-based metabolomics. *Molecular bioSystems*, 8(2):470–81, February 2012.
- [38] Monya Baker. Metabolomics: from small molecules to big ideas. *Nature Methods*, 8(2):117–121, February 2011.
- [39] Susan D Richardson. Environmental mass spectrometry: emerging contaminants and current issues. *Analytical chemistry*, 82(12):4742–74, June 2010.
- [40] J William Allwood and Royston Goodacre. An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical analysis : PCA*, 21(1):33–47, 2010.
- [41] Ken Webb, Tony Bristow, and Mike Sargent. Methodology for Accurate Mass Measurement of Small Molecules Best Practice Guide. Technical report, LGC Ltd., 2004.
- [42] Maud M. Koek, Renger H. Jellema, Jan van der Greef, Albert C. Tas, and Thomas Hankemeier. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics : Official journal of the Metabolomic Society*, 7(3):307–328, September 2011.
- [43] Liang Cui, Yie Hou Lee, Yadunanda Kumar, Fengguo Xu, Kun Lu, Eng Eong Ooi, Steven R. Tannenbaum, and Choon Nam Ong. Serum Metabolome and Lipidome Changes in Adult Patients with Primary Dengue Infection. *PLoS Neglected Tropical Diseases*, 7(8):e2373, August 2013.
- [44] Carsten Denkert, Jan Budczies, Tobias Kind, Wilko Weichert, Peter Tablack, Jalid Sehouli, Silvia Niesporek, Dominique Könsgen, Manfred Dietel, and Oliver Fiehn. Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. *Cancer research*, 66(22):10795–804, November 2006.
- [45] Masahiro Sugimoto, David T Wong, Akiyoshi Hirayama, Tomoyoshi Soga, and Masaru Tomita. Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. *Metabolomics : Official journal of the Metabolomic Society*, 6(1):78–95, March 2010.
- [46] Ghaniah Hassan-Smith, Graham R Wallace, Michael R Douglas, and Alexandra J Sinclair. The role of metabolomics in neurological disease. *Journal of neuroimmunology*, 248(1-2):48–52, July 2012.
- [47] Elizabeth J Want, Ian D Wilson, Helen Gika, Georgios Theodoridis, Robert S Plumb, John Shockcor, Elaine Holmes, and Jeremy K Nicholson. Global metabolic profiling procedures for urine using UPLC-MS. *Nature protocols*, 5(6):1005–18, January 2010.
- [48] Xinjie Zhao, Jens Fritsche, Jiangshan Wang, Jing Chen, Kilian Rittig, Philippe Schmitt-Kopplin, Andreas Fritsche, Hans-Ulrich Häring, Erwin D. Schleicher, Guowang Xu, and Rainer Lehmann. Metabonomic fingerprints of fasting plasma and spot urine reveal human pre-diabetic metabolic traits. *Metabolomics*, 6(3):362–374, March 2010.
- [49] Anna Floegel, Norbert Stefan, Zhonghao Yu, Kristin Mühlenbruch, Dagmar Drogan, Hans-Georg Joost, Andreas Fritsche, Hans-Ulrich Häring, Martin Hrabě de Angelis, Annette Peters, Michael Roden, Cornelia Prehn, Rui Wang-Sattler, Thomas Illig, Matthias B Schulze, Jerzy Adamski, Heiner Boeing, and Tobias Pischon. Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes*, 62(2):639–48, February 2013.
- [50] Fatima a. Nasrallah, Vladimir J. Balcar, and Caroline Rae. A metabonomic study of inhibition of GABA uptake in the cerebral cortex. *Metabolomics*, 6(1):67–77, August 2009.
- [51] Rima Kaddurah-Daouk, Rebecca A Baillie, Hongjie Zhu, Zhao-Bang Zeng, Michelle M Wiest, Uyen Thao Nguyen, Steven M Watkins, and Ronald M Krauss. Lipidomic analysis of variation in response to simvastatin in the Cholesterol and Pharmacogenetics Study. *Metabolomics : Official journal of the Metabolomic Society*, 6(2):191–201, June 2010.

## REFERENCES

---

- [52] Susan Schiavo Bird, Vasant R Marur, Matthew J Sniatynski, Heather K Greenberg, and Bruce S Kristal. Serum lipidomics profiling using LC-MS and high-energy collisional dissociation fragmentation: focus on triglyceride detection and characterization. *Analytical chemistry*, 83(17):6648–57, September 2011.
- [53] Todd R Sandrin, Jason E Goldstein, and Stephanie Schumaker. MALDI TOF MS profiling of bacteria at the strain level: a review. *Mass spectrometry reviews*, 32(3):188–217, 2012.
- [54] Jean-Philippe Lavigne, Paula Espinal, Catherine Dunyach-Remy, Nourredine Messad, Alix Pantel, and Albert Sotto. Mass spectrometry: a revolution in clinical microbiology? *Clinical chemistry and laboratory medicine : CCLM / FESCC*, pages 1–14, October 2012.
- [55] M. Wink. Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theoretical and Applied Genetics*, 75(2):225–233, January 1988.
- [56] O Fiehn, J Kopka, P Dörmann, T Altmann, R N Trethewey, and L Willmitzer. Metabolite profiling for plant functional genomics. *Nature biotechnology*, 18(11):1157–61, November 2000.
- [57] Patricia M Cano, Emilien L Jamin, Souria Tadrst, Pascal Bourdaud’hui, Michel Péan, Laurent Debrauwer, Isabelle P Oswald, Marcel Delaforge, and Olivier Puel. New untargeted metabolic profiling combining mass spectrometry and isotopic labeling: application on *Aspergillus fumigatus* grown on wheat. *Analytical chemistry*, 85(17):8412–20, September 2013.
- [58] Ric C H De Vos, Sofia Moco, Arjen Lommen, Joost J B Keurentjes, Raoul J Bino, and Robert D Hall. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature protocols*, 2(4):778–91, January 2007.
- [59] Mark R. Viant and Ulf Sommer. Mass spectrometry based environmental metabolomics: a primer and review. *Metabolomics*, 9(S1):144–158, March 2012.
- [60] Atsushi Ishihara, Fumio Matsuda, Hisashi Miyagawa, and Kyo Wakasa. Metabolomics for metabolically manipulated plants: effects of tryptophan overproduction. *Metabolomics*, 3(3):319–334, September 2007.
- [61] Yaniv Semel, Nicolas Schauer, Ute Roessner, Dani Zamir, and Alisdair Robert Fernie. Metabolite analysis for the comparison of irrigated and non-irrigated field grown tomato of varying genotype. *Metabolomics*, 3(3):289–295, June 2007.
- [62] Georgios Theodoridis, Helen Gika, Pietro Franceschi, Lorenzo Caputi, Panagiotis Arapitsas, Mattias Scholz, Domenico Masuero, Ron Wehrens, Urska Vrhovsek, and Fulvio Mattivi. LC-MS based global metabolite profiling of grapes: solvent extraction protocol optimisation. *Metabolomics*, 8(2):175–185, March 2011.
- [63] Yozo Okazaki, Yukiko Kamide, Masami Yokota Hirai, and Kazuki Saito. Plant lipidomics based on hydrophilic interaction chromatography coupled to ion trap time-of-flight mass spectrometry. *Metabolomics : Official journal of the Metabolomic Society*, 9(Suppl 1):121–131, March 2013.
- [64] Jiangjiang Liu, He Wang, R Graham Cooks, and Zheng Ouyang. Leaf spray: direct chemical analysis of plant material and living plants by mass spectrometry. *Analytical chemistry*, 83(20):7608–13, October 2011.
- [65] Peter W Carr, Dwight R Stoll, and Xiaoli Wang. Perspectives on recent advances in the speed of high-performance liquid chromatography. *Analytical chemistry*, 83(6):1890–900, March 2011.
- [66] Tania Portoles, Elena Pitarch, Francisco J Lopez, Felix Hernandez, and Wilfried M A Niessen. Use of soft and hard ionization techniques for elucidation of unknown compounds by gas chromatography/time-of-flight mass spectrometry. *Data Processing*, 25:1589–1599, 2011.
- [67] Bhagwat Prasad, Amit Garg, Hardik Takwani, and Saranjit Singh. Metabolite identification by liquid chromatography-mass spectrometry. *TrAC Trends in Analytical Chemistry*, 30(2):360–387, February 2011.
- [68] Sara Forcisi, Franco Moritz, Basem Kanawati, Dimitrios Tziotis, Rainer Lehmann, and Philippe Schmitt-Kopplin. Liquid chromatography-mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *Journal of chromatography. A*, 1292:51–65, April 2013.
- [69] John Draper, Amanda J. Lloyd, Royston Goodacre, and Manfred Beckmann. Flow infusion electrospray ionisation mass spectrometry for high throughput, non-targeted metabolite fingerprinting: a review. *Metabolomics*, 9(S1):4–29, July 2012.
- [70] Jürgen Hartler, Ravi Tharakan, Harald C Köfeler, David R Graham, and Gerhard G Thallinger. Bioinformatics tools and challenges in structural analysis of lipidomics MS/MS data. *Briefings in bioinformatics*, 14(3):375–90, May 2013.



## REFERENCES

---

- [71] David G Watson. A Rough Guide to Metabolite Identification Using High Resolution Liquid Chromatography Mass Spectrometry in Metabolomic Profiling in Metazoans. *Computational and structural biotechnology journal*, 4(January):e201301005, January 2013.
- [72] Phil Price. Standard definitions of terms relating to mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 2(4):336–348, August 1991.
- [73] Ryo Taguchi, Mashahiro Nishijima, and Takao Shimizu. Basic analytical systems for lipidomics by mass spectrometry in Japan. *Methods in Enzymology*, 432:185–211, 2007.
- [74] Ambrose Furey, Merisa Moriarty, Vaishali Bane, Brian Kinsella, and Mary Lehane. Ion suppression; A critical review on causes, evaluation, prevention and applications. *Talanta*, 115:104–22, October 2013.
- [75] Thomas M Annesley. Ion suppression in mass spectrometry. *Clinical Chemistry*, 49:1041–1044, 2003.
- [76] WH McFadden and EA Day. Scan Rate Considerations in Combined Gas Chromatography-Mass Spectrometry. *Analytical Chemistry*, 36(12):2362–2363, 1964.
- [77] IUPAC. *IUPAC Compendium of Chemical Terminology*. IUPAC, Research Triangle Park, NC, June 2009. ISBN 0-9678550-9-8.
- [78] Michael P Balogh and Waters Corp. Debating Resolution and Mass Accuracy. *Journal of the American Society for Mass Spectrometry*, 17(3), 2004.
- [79] Gaetan Glauser, Nathalie Veyrat, Bertrand Rochat, Jean-Luc Wolfender, and Ted C J Turlings. Ultra-high pressure liquid chromatography-mass spectrometry for plant metabolomics: a systematic comparison of high-resolution quadrupole-time-of-flight and single stage Orbitrap mass spectrometers. *Journal of chromatography. A*, 1292:151–9, May 2013.
- [80] Jeonghoon Lee and Peter T a Reilly. Limitation of Time-of-Flight Resolution in the Ultra High Mass Range. *Analytical chemistry*, pages 5–7, July 2011.
- [81] Warwick B Dunn. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology*, 5:011001, 2008.
- [82] Seongho Kim, Imhoi Koo, Aiqin Fang, and Xiang Zhang. Smith-Waterman peak alignment for comprehensive two-dimensional gas chromatography-mass spectrometry. *BMC bioinformatics*, 12(1):235, January 2011.
- [83] Miquel Rojas-Cherto, Julio E Peironcely, Piotr T Kasper, Justin J J van der Hooft, Ric C H de Vos, Rob Vreeken, Thomas Hankemeier, and Theo Reijmers. Metabolite Identification Using Automated Comparison of High-Resolution Multistage Mass Spectral Trees. *Analytical chemistry*, June 2012.
- [84] Stephen Stein. Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Analytical chemistry*, 84(17):7274–82, September 2012.
- [85] Jiyang Zhang, Jie Ma, Wei Zhang, Changming Xu, Yunping Zhu, and Hongwei Xie. *FTDR 2.0: a tool to achieve sub-ppm level recalibrated accuracy in routine LC-MS analysis*. July 2013. ISBN 8673184576.
- [86] Helen G Gika, Georgios a Theodoridis, Robert S Plumb, and Ian D Wilson. Current practice of liquid chromatography-mass spectrometry in metabolomics and metabonomics. *Journal of pharmaceutical and biomedical analysis*, 87:12–25, January 2014.
- [87] Igor Nikolskiy, Nathaniel G Mahieu, Ying-Jr Chen, Ralf Tautenhahn, and Gary J Patti. An untargeted metabolomic workflow to improve structural characterization of metabolites. *Analytical chemistry*, 85(16):7713–9, August 2013.
- [88] Tairo Ogura, Takeshi Bamba, and Eiichiro Fukusaki. Development of a practical metabolite identification technique for non-targeted metabolomics. *Journal of Chromatography A*, pages 1–7, May 2013.
- [89] Esther SI Chong, Tony K McGhie, Julian a Heyes, and Kathryn M Stowell. Metabolite profiling and quantification of phytochemicals in potato extracts using ultra high performance liquid chromatography-mass spectrometry. *Journal of the science of food and agriculture*, June 2013.
- [90] Julio E Peironcely, Miguel Rojas-Chertó, Albert Tas, Rob Vreeken, Theo Reijmers, Leon Coulier, and Thomas Hankemeier. Automated pipeline for de novo metabolite identification using mass-spectrometry-based metabolomics. *Analytical chemistry*, 85(7):3576–83, April 2013.
- [91] Lars J Kangas, Thomas O Metz, Giorgis Isaac, Brian T Schrom, Bojana Ginovska-Pangovska, Luning Wang, Li Tan, Robert R Lewis, and John H Miller. In silico identification software (ISIS): a machine

## REFERENCES

---

- learning approach to tandem mass spectral identification of lipids. *Bioinformatics (Oxford, England)*, 28(13):1705–13, July 2012.
- [92] Karen W Phinney, Guillaume Ballihaut, Mary Bedner, Brandi S Benford, Johanna E Camara, Steven J Christopher, W Clay Davis, Nathan G Dodder, Gauthier Eppe, Brian E Lang, Stephen E Long, Mark S Lowenthal, Elizabeth a McGaw, Karen E Murphy, Bryant C Nelson, Jocelyn L Prendergast, Jessica L Reiner, Catherine a Rimmer, Lane C Sander, Michele M Schantz, Katherine E Sharpless, Lorna T Sniegowski, Susan S-C Tai, Jeanice B Thomas, Thomas W Vetter, Michael J Welch, Stephen a Wise, Laura J Wood, William F Guthrie, Charles R Hagwood, Stefan D Leigh, James H Yen, Nien-Fan Zhang, Madhu Chaudhary-Webb, Huiping Chen, Zia Fazili, Donna J LaVoie, Leslie F McCoy, Shahzad S Momin, Neelima Paladugula, Elizabeth C Pendergrast, Christine M Pfeiffer, Carissa D Powers, Daniel Rabinowitz, Michael E Rybak, Rosemary L Schleicher, Bridgette M H Toombs, Mary Xu, Mindy Zhang, and Arthur L Castle. Development of a Standard Reference Material for metabolomics research. *Analytical chemistry*, 85(24):11732–8, December 2013.
- [93] Marie Brown, Warwick B. Dunn, David I. Ellis, Royston Goodacre, Julia Handl, Joshua D. Knowles, Steve OHagan, Irena Spasić, and Douglas B. Kell. A metabolome pipeline: from concept to data to knowledge. *Metabolomics*, 1(1):39–51, March 2005.
- [94] Patrick G a Pedrioli, Jimmy K Eng, Robert Hubley, Mathijs Vogelzang, Eric W Deutsch, Brian Raught, Brian Pratt, Erik Nilsson, Ruth H Angeletti, Rolf Apweiler, Kei Cheung, Catherine E Costello, Henning Hermjakob, Sequin Huang, Randall K Julian, Eugene Kapp, Mark E McComb, Stephen G Oliver, Gilbert Omenn, Norman W Paton, Richard Simpson, Richard Smith, Chris F Taylor, Weimin Zhu, and Ruedi Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, 22(11):1459–66, November 2004.
- [95] Sandra Orchard, Luisa Montechi-Palazzi, Eric W Deutsch, Pierre-Alain Binz, Andrew R Jones, Norman Paton, Angel Pizarro, David M Creasy, Jérôme Wojcik, and Henning Hermjakob. 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):3436–40, October 2007.
- [96] E. Deutsch. mzML: a single, unifying data format for mass spectrometer output. *Proteomics*, 8(14):2776–2777, 2008.
- [97] Eric W Deutsch. File formats commonly used in mass spectrometry proteomics. *Molecular & cellular proteomics : MCP*, (1):1–32, September 2012.
- [98] R Rew and G Davis. NetCDF: an interface for scientific data access. *IEEE Computer Graphics and Applications*, 10:76–82, 1990.
- [99] Sandra Orchard. Data standardization and sharing-the work of the HUPO-PSI. *Biochimica et biophysica acta*, 1844(1 Pt A):82–7, January 2014.
- [100] Julian L Griffin and Christoph Steinbeck. So what have data standards ever done for us? The view from metabolomics. *Genome medicine*, 2(6):38, January 2010.
- [101] Shaji Krishnan, Jack T W E Vogels, Leon Coulier, Richard C Bas, Margriet W B Hendriks, Thomas Hankemeier, and Uwe Thissen. Instrument and process independent binning and baseline correction methods for liquid chromatography-high resolution-mass spectrometry deconvolution. *Analytica chimica acta*, 740:12–9, August 2012.
- [102] Peijuan Zhu, Wei Ding, Wei Tong, Anima Ghosal, Kevin Alton, and Swapan Chowdhury. A retention-time-shift-tolerant background subtraction and noise reduction algorithm (BgS-NoRA) for extraction of drug metabolites in liquid chromatography/mass spectrometry data from biological matrices. *Rapid communications in mass spectrometry : RCM*, 23(11):1563–72, June 2009.
- [103] Haiying Zhang and Yanou Yang. An algorithm for thorough background subtraction from high-resolution LC/MS data: application for detection of glutathione-trapped reactive metabolites. *Journal of mass spectrometry*, 43(9):1181–1190, 2008.
- [104] Zhanfeng Xu, Xiaobo Sun, and Peter De B Harrington. Baseline correction method using an orthogonal basis for gas chromatography/mass spectrometry data. *Analytical chemistry*, 83(19):7464–71, October 2011.
- [105] Kirill A Veselkov, Lisa K Vingara, Perrine Masson, Steven L Robinette, Elizabeth Want, Jia V Li, Richard H Barton, Claire Boursier-Neyret, Bernard Walther, Timothy M Ebbels, István Pelczar, Elaine Holmes, John C Lindon, and Jeremy K Nicholson. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical chemistry*, 83(15):5864–72, August 2011.

## REFERENCES

---

- [106] Willem Windig, J. Martin Phalp, and Alan W. Payne. A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry*, 68(20):3602–3606, January 1996.
- [107] W Windig. The use of the Durbin-Watson criterion for noise and background reduction of complex liquid chromatography/mass spectrometry data and a new algorithm to determine sample differences. *Chemometrics and Intelligent Laboratory Systems*, 77:206 – 214, December 2004.
- [108] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2):89–121, May 1996.
- [109] J Luo, K Ying, and J Bai. SavitzkyGolay smoothing and differentiation filter for even number data. *Signal Processing*, 85:1429–1434, 2005.
- [110] Bernd O Keller, Jie Sui, Alex B Young, and Randy M Whittal. Interferences and contaminants encountered in modern mass spectrometry. *Analytica chimica acta*, 627(1):71–81, October 2008.
- [111] Xiaoli Wei, Xue Shi, Seongho Kim, Li Zhang, Jeffrey S Patrick, Joe Binkley, Craig McClain, and Xiang Zhang. Data preprocessing method for liquid chromatography-mass spectrometry based metabolomics. *Analytical chemistry*, 84(18):7963–71, September 2012.
- [112] Jianqiu Zhang, Elias Gonzalez, Travis Hestilow, William Haskins, and Yufei Huang. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Current genomics*, 10(6):388–401, September 2009.
- [113] S.E. Stein. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, 10(8):770–781, August 1999.
- [114] B.N. Colby. Spectral deconvolution for overlapping GC/MS components. *Journal of the American Society for Mass Spectrometry*, 3(5):558–562, 1992.
- [115] Miso Project for the creation of high-quality interactive storytelling and data visualisation content.
- [116] Tianwei Yu and Hesen Peng. Quantification and deconvolution of asymmetric LC-MS peaks using the bi-Gaussian mixture model and statistical model selection. *BMC bioinformatics*, 11(1):559, November 2010.
- [117] Renger H. Jellema, Shaji Krishnan, Margriet M.W.B. Hendriks, Bas Muilwijk, and Jack T.W.E. Vogels. Deconvolution using signal segmentation. *Chemometrics and Intelligent Laboratory Systems*, 104(1):132–139, November 2010.
- [118] J Jiang. Principles and methodologies in self-modeling curve resolution. *Chemometrics and Intelligent Laboratory Systems*, 71(1):1–12, April 2004.
- [119] Pan Du, Warren a Kibbe, and Simon M Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics (Oxford, England)*, 22(17):2059–65, September 2006.
- [120] Bernd Fischer, Jonas Grossmann, Volker Roth, Wilhelm Gruissem, Sacha Baginsky, and Joachim M Buhmann. Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics (Oxford, England)*, 22(14):e132–40, July 2006.
- [121] John T Prince and Edward M Marcotte. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical chemistry*, 78(17):6140–52, September 2006.
- [122] Q Peter He, Jin Wang, James a Mobley, Joshua Richman, and William E Grizzle. Self-calibrated warping for mass spectra alignment. *Cancer informatics*, 10:65–82, January 2011.
- [123] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11:561–580, 2007.
- [124] Venura Perera, Marta Torres Zabala, Hannah Florance, Nicholas Smirnoff, Murray Grant, and Zheng Rong Yang. Aligning extracted LC-MS peak lists via density maximization. *Metabolomics*, 8(S1):175–185, December 2011.
- [125] Rudolf Fruehwirth, D. R. Mani, and Saumyadipta Pyne. Clustering with position-specific constraints on variance: Applying redescending M-estimators to label-free LC-MS data analysis. *BMC Bioinformatics*, 12(1):358, 2011.

## REFERENCES

---

- [126] Radka Stoyanova, Andrew W Nicholls, Jeremy K Nicholson, John C Lindon, and Truman R Brown. Automatic alignment of individual peaks in large high-resolution spectral data sets. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 170(2):329–35, October 2004.
- [127] Frans M van der Kloet, Margriet Hendriks, Thomas Hankemeier, and Theo Reijmers. A new approach to untargeted integration of high resolution liquid chromatography-mass spectrometry data. *Analytica chimica acta*, 801:34–42, November 2013.
- [128] Y M Tikunov, S Laptinok, R D Hall, A Bovy, and R C H de Vos. MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics : Official journal of the Metabolomic Society*, 8(4):714–718, August 2012.
- [129] Carsten Kuhl, Ralf Tautenhahn, Christoph Böttcher, Tony R Larson, and Steffen Neumann. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical chemistry*, 84(1):283–9, January 2012.
- [130] Shaji Krishnan, Elwin E R Verheij, Richard C Bas, Margriet W B Hendriks, Thomas Hankemeier, Uwe Thissen, and Leon Coulier. Pre-processing liquid chromatography/high-resolution mass spectrometry data: extracting pure mass spectra by deconvolution from the invariance of isotopic distribution. *Rapid communications in mass spectrometry : RCM*, 27(9):917–923, May 2013.
- [131] Masahiro Sugimoto, Masato Kawakami, Martin Robert, Tomoyoshi Soga, and Masaru Tomita. Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. *Current bioinformatics*, 7(1):96–108, March 2012.
- [132] Tytus D Mak, Evagelia C Laiakis, Maryam Goudarzi, and Albert J Fornace. MetaboLyzer: a novel statistical workflow for analyzing Postprocessed LC-MS metabolomics data. *Analytical chemistry*, 86(1):506–13, January 2014.
- [133] Ai-Hua Zhang, Hui Sun, Ying Han, Guang-Li Yan, Ye Yuan, Gao-Chen Song, Xiao-Xia Yuan, Ning Xie, and Xi-Jun Wang. Ultraperformance liquid chromatography-mass spectrometry based comprehensive metabolomics combined with pattern recognition and network analysis methods for characterization of metabolites and metabolic pathways from biological data sets. *Analytical chemistry*, 85(15):7606–12, August 2013.
- [134] Barry K Lavine and Jerome Workman. Chemometrics. *Analytical chemistry*, 85(2):705–14, January 2013.
- [135] Robert a van den Berg, Huub C J Hoefsloot, Johan a Westerhuis, Age K Smilde, and Mariët J van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7:142, January 2006.
- [136] Grant Hughes, Charmion Cruickshank-Quinn, Richard Reisdorph, Sharon Lutz, Irina Petrache, Nichole Reisdorph, Russell Bowler, and Katerina Kechris. MSPrep–Summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data. *Bioinformatics (Oxford, England)*, 30(1):133–134, November 2013.
- [137] Charles R. Warren. Use of chemical ionization for GCMS metabolite profiling. *Metabolomics*, 9(S1):110–120, August 2011.
- [138] Martin a Ott and Gert Vriend. Correcting ligands, metabolites, and pathways. *BMC bioinformatics*, 7:517, January 2006.
- [139] Royston Goodacre, David Broadhurst, Age K. Smilde, Bruce S. Kristal, J. David Baker, Richard Beger, Conrad Bessant, Susan Connor, Giorgio Capuani, Andrew Craig, Tim Ebbels, Douglas B. Kell, Cesare Manetti, Jack Newton, Giovanni Paternostro, Ray Somorjai, Michael Sjöström, Johan Trygg, and Florian Wulfert. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3):231–241, August 2007.
- [140] Denis V. Rubtsov, Helen Jenkins, Christian Ludwig, John Easton, Mark R. Viant, Ulrich Günther, Julian L. Griffin, and Nigel Hardy. Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics*, 3(3):223–229, September 2007.
- [141] Reza M Salek, Christoph Steinbeck, Mark R Viant, Royston Goodacre, and Warwick B Dunn. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience*, 2(1):13, October 2013.
- [142] NIH. The Metabolomics Data Center and Workbench (MDCW).
- [143] Christoph Steinbeck, Pablo Conesa, Kenneth Haug, Tejasvi Mahendraker, Mark Williams, Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta Sansone, Reza M. Salek, and Julian L. Griffin. Metabo-

## REFERENCES

---

- Lights: towards a new COSMOS of metabolomics data management. *Metabolomics : Official journal of the Metabolomic Society*, 8(5):757–760, October 2012.
- [144] Raphael Aggio, Silas Granato Villas-Bôas, and Katya Ruggiero. Metab: an R package for high-throughput analysis of metabolomics data generated by GC-MS. *Bioinformatics (Oxford, England)*, 27(16):2316–8, August 2011.
- [145] Jianguo Xia, Nick Psychogios, Nelson Young, and David S Wishart. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research*, 37(Web Server issue):W652–60, July 2009.
- [146] Xiaoli Wei, Wenlong Sun, Xue Shi, Imhoi Koo, Bing Wang, Jun Zhang, Xinmin Yin, Yunan Tang, Bogdan Bogdanov, Seongho Kim, Zhanxiang Zhou, Craig McClain, and Xiang Zhang. MetSign: a computational platform for high-resolution mass spectrometry-based metabolomics. *Analytical chemistry*, 83(20):7668–75, October 2011.
- [147] Tomás Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Oresic. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, 11:395, January 2010.
- [148] So Jeong Yun, Ji-Won Park, Il Ju Choi, Byeongsoo Kang, Hark Kyun Kim, Dae Won Moon, Tae Geol Lee, and Daehee Hwang. TOFSIMS-P: A web-based platform for analysis of large-scale TOF-SIMS data. *Analytical chemistry*, November 2011.
- [149] Sean O’Callaghan, David P De Souza, Andrew Isaac, Qiao Wang, Luke Hodkinson, Moshe Olshansky, Tim Erwin, Bill Appelbe, Dedreia L Tull, Ute Roessner, Antony Bacic, Malcolm J McConville, and Vladimir A Likić. PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools. *BMC bioinformatics*, 13:115, January 2012.
- [150] A Bertsch, C Gröpl, K Reinert, and O Kohlbacher. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods in molecular biology (Clifton, N.J.)*, 696:353–367, 2011.
- [151] Patrick Kiefer, Uwe Schmitt, and Julia a Vorholt. eMZed: an open source framework in Python for rapid and interactive development of LC/MS data analysis workflows. *Bioinformatics (Oxford, England)*, 29(7):963–4, April 2013.
- [152] Zhentian Lei, Haiquan Li, Junil Chang, Patrick X. Zhao, and Lloyd W. Sumner. MET-IDEA version 2.06; improved efficiency and additional functions for mass spectrometry-based metabolomics data processing. *Metabolomics*, February 2012.
- [153] Karsten Hiller, Jasper Hangebrauk, Christian Jäger, Jana Spura, Kerstin Schreiber, and Dietmar Schomburg. MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Analytical chemistry*, 81(9):3429–39, May 2009.
- [154] Eugene Melamud, Livia Vastag, and Joshua D Rabinowitz. Metabolomic analysis and visualization engine for LC-MS data. *Analytical chemistry*, 82(23):9818–26, December 2010.
- [155] Andris Jankevics, Maria Elena Merlo, Marcel de Vries, Roel J. Vonk, Eriko Takano, and Rainer Breitling. Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets. *Metabolomics : Official journal of the Metabolomic Society*, 8(Suppl 1):29–36, June 2012.
- [156] Erik Tengstrand, Johan Lindberg, and K Magnus Aberg. TracMass 2 - a modular suite of tools for processing chromatography-full scan mass spectrometry data. *Analytical chemistry*, March 2014.
- [157] Francesc Fernández-Albert, Rafael Llorach, Cristina Andrés-Lacueva, and Alexandre Perera. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). *Bioinformatics (Oxford, England)*, pages 2–3, March 2014.
- [158] Colin a Smith, Elizabeth J Want, Grace O’Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, 78(3):779–87, February 2006.
- [159] Arjen Lommen. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Analytical chemistry*, 81(8):3079–86, April 2009.
- [160] Ralf Tautenhahn, Gary J Patti, Duane Rinehart, and Gary Siuzdak. XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical chemistry*, 84(11):5035–9, June 2012.
- [161] Alvaro Cuadros-Inostroza, Camila Caldana, Henning Redestig, Miyako Kusano, Jan Lisec, Hugo Peña Cortés, Lothar Willmitzer, and Matthew a Hannah. TargetSearch—a Bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data. *BMC bioinformatics*, 10:428, January 2009.

## REFERENCES

---

- [162] Yuliya V Karpievitch, Elizabeth G Hill, Adam J Smolka, Jeffrey S Morris, Kevin R Coombes, Keith a Baggerly, and Jonas S Almeida. PrepMS: TOF MS data graphical preprocessing tool. *Bioinformatics (Oxford, England)*, 23(2):264–5, January 2007.
- [163] Hiroki Takahashi, Takuya Morimoto, Naotake Ogasawara, and Shigehiko Kanaya. AMDORAP: Non-targeted metabolic profiling based on high-resolution LC-MS. *BMC Bioinformatics*, 12(1):259, 2011.
- [164] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, Giuseppe Tradigo, and Pierangelo Veltri. MaSDA: a system for analyzing mass spectrometry data. *Computer methods and programs in biomedicine*, 95(2 Suppl):S12–21, August 2009.
- [165] Xiaojing Huang, Ying-Jr Chen, Kevin Cho, Igor Nikolskiy, Peter A Crawford, and Gary J Patti. X13CMS: global tracking of isotopic labels in untargeted metabolomics. *Analytical chemistry*, 86(3):1632–9, February 2014.
- [166] Florence Nicolè, Yann Guitton, Elodie A Courtois, Sandrine Moja, Laurent Legendre, and Martine Hossaert-McKey. MSeasy: unsupervised and untargeted GC-MS data processing. *Bioinformatics (Oxford, England)*, 28(17):2278–80, September 2012.
- [167] Lochana C Menikarachchi, Shannon Cawley, Dennis W Hill, L Mark Hall, Lowell Hall, Steven Lai, Janine Wilder, and David F Grant. MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures. *Analytical chemistry*, 84(21):9388–94, November 2012.
- [168] John Draper, David P Enot, David Parker, Manfred Beckmann, Stuart Snowdon, Wanchang Lin, and Hassan Zubair. Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC bioinformatics*, 10:227, January 2009.
- [169] Michael Gerlich and Steffen Neumann. MetFusion: integration of compound identification strategies. *Journal of mass spectrometry : JMS*, 48(3):291–8, March 2013.
- [170] Miguel Rojas-Chertó, Michael van Vliet, Julio E Peironcely, Ronnie van Doorn, Maarten Kooyman, Tim te Beek, Marc a van Driel, Thomas Hankemeier, and Theo Reijmers. MetiTree: a web application to organize and process high-resolution multi-stage mass spectrometry metabolomics data. *Bioinformatics (Oxford, England)*, 28(20):2707–9, October 2012.
- [171] Julio E Peironcely, Miguel Rojas-Chertó, Albert Tas, Rob Vreeken, Theo Reijmers, Leon Coulier, and Thomas Hankemeier. Automated pipeline for de novo metabolite identification using mass-spectrometry-based metabolomics. *Analytical chemistry*, 85(7):3576–83, April 2013.
- [172] Igor Nikolskiy, Nathaniel G Mahieu, Ying-Jr Chen, Ralf Tautenhahn, and Gary Joseph Patti. An untargeted metabolomic workflow to improve structural characterization of metabolites. *Analytical chemistry*, 85(16):7713–9, August 2013.
- [173] Robert Cho, Yingying Huang, Jae C. Schwartz, Yan Chen, Timothy J. Carlson, and Ji Ma. MS(M), an efficient workflow for metabolite identification using hybrid linear ion trap Orbitrap mass spectrometer. *Journal of the American Society for Mass Spectrometry*, 23(5):880–8, May 2012.
- [174] Steffen Neumann, Andrea Thum, and Christoph Böttcher. Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. *Metabolomics*, 9(S1):84–91, February 2012.
- [175] Mikko Katajamaa and Matej Oresic. Data processing for mass spectrometry-based metabolomics. *Journal of chromatography. A*, 1158(1-2):318–28, July 2007.
- [176] Perttu Haimi, Andreas Uphoff, Martin Hermansson, and Pentti Somerharju. Software tools for analysis of mass spectrometric lipidome data. *Analytical chemistry*, 78(24):8324–31, December 2006.
- [177] Vivien Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, June 2013.
- [178] Seongho Kim and Xiang Zhang. Comparative analysis of mass spectral similarity measures on peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry. *Computational and mathematical methods in medicine*, 2013:509761, January 2013.
- [179] Arjen Lommen, Henk J van der Kamp, Harrie J Kools, Martijn K van der Lee, Guido van der Weg, and Hans G J Mol. metAlignID: A high-throughput software tool set for automated detection of trace level contaminants in comprehensive LECO two-dimensional gas chromatography time-of-flight mass spectrometry data. *Journal of chromatography. A*, September 2012.
- [180] Cheolhwan Oh, Xiaodong Huang, Fred E Regnier, Charles Buck, and Xiang Zhang. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry peak sorting algorithm. *Journal of chromatography. A*, 1179(2):205–15, February 2008.

## REFERENCES

---

- [181] Christos A Ouzounis. Rise and demise of bioinformatics? Promise and progress. *PLoS computational biology*, 8(4):e1002487, January 2012.
- [182] Peter Willett. Chemoinformatics: A history. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(1):46–56, January 2011.
- [183] David B. Searls. The Roots of Bioinformatics. *PLoS Computational Biology*, 6(6):e1000809, June 2010.
- [184] Christopher M Dobson. Chemical space and biology. *Nature*, 432(7019):824–8, December 2004.
- [185] Evan E Bolton, Yanli Wang, Paul A Thiessen, and Stephen H Bryant. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry*, 4, 2008.
- [186] Harry E Pence and Antony Williams. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education*, 87:1123–1124, 2010.
- [187] Joerg Kurt Wegner, Egon Willighagen, Aaron Sterling, Rajarshi Guha, Andreas Bender, Jean-Loup Faulon, Janna Hastings, Noel O’Boyle, John Overington, and Herman Van Vlijmen. Cheminformatics. *Communications of the ACM*, 55(11):65, November 2012.
- [188] Egon L Willighagen, Noel M O’Boyle, Harini Gopalakrishnan, Dazhi Jiao, Rajarshi Guha, Christoph Steinbeck, and David J Wild. Userscripts for the life sciences. *BMC bioinformatics*, 8:487, January 2007.
- [189] Luc Patiny and Alain Borel. ChemCalc: a building block for tomorrow’s chemical infrastructure. *Journal of chemical information and modeling*, 53(5):1223–8, May 2013.
- [190] Stefan Höck and Rainer Riedl. chemf: A purely functional chemistry toolkit. *Journal of cheminformatics*, 4(1):38, December 2012.
- [191] Greg Landrum. RDKit.
- [192] Christoph Steinbeck, Christian Hoppe, Stefan Kuhn, Matteo Floris, Rajarshi Guha, and Egon L Willighagen. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Current pharmaceutical design*, 12(17):2111–20, January 2006.
- [193] Noel M O’Boyle, Michael Banck, Craig a James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, October 2011.
- [194] D. Fourches, E. Muratov, and A. Tropsha. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of chemical information and modeling*, 50(7):1189–1204, 2010.
- [195] Kevin R Lawson and Jonty Lawson. LICSS - A chemical spreadsheet in Microsoft Excel. *Journal of cheminformatics*, 4(1):3, February 2012.
- [196] Daniel M Lowe, Peter T Corbett, Peter Murray-Rust, and Robert C Glen. Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of chemical information and modeling*, 51(3):739–53, March 2011.
- [197] Jon Chambers, Mark Davies, Anna Gaulton, Anne Hersey, Sameer Velankar, Robert Petryszak, Janna Hastings, Louisa Bellis, Shaun McGlinchey, and John P Overington. UniChem: a unified chemical structure cross-referencing and identifier tracking system. *Journal of cheminformatics*, 5(1):3, January 2013.
- [198] Ola Spjuth, Jonathan Alvarsson, Arvid Berg, Martin Eklund, Stefan Kuhn, Carl Mäsak, Gilleain Torrance, Johannes Wagener, Egon L Willighagen, Christoph Steinbeck, and Jarl E S Wikberg. Bioclipse 2: a scriptable integration platform for the life sciences. *BMC bioinformatics*, 10:397, January 2009.
- [199] Matthias Hilbig, Sascha Urbaczek, Inken Groth, Stefan Heuser, and Matthias Rarey. MONA - Interactive manipulation of molecule collections. *Journal of Cheminformatics*, 5(1):38, 2013.
- [200] WA Warr. Representation of chemical structures. *Wiley Interdisciplinary Reviews: Computational*, 1 (August):557–579, 2011.
- [201] William J. Wiswesser. How the WLN began in 1949 and how it might be in 1999. *Journal of Chemical Information and Modeling*, 22(2):88–93, May 1982.
- [202] Jacques Emile Dubois and Yves Sobel. DARC system for documentation and artificial intelligence in chemistry. *Journal of Chemical Information and Modeling*, 25(3):326–333, August 1985.
- [203] Rajarshi Guha, Michael T Howard, Geoffrey R Hutchison, Peter Murray-Rust, Henry Rzepa, Christoph Steinbeck, Jörg Wegner, and Egon L Willighagen. The Blue Obelisk-interoperability in chemical informatics. *Journal of chemical information and modeling*, 46(3):991–8, 2006.

## REFERENCES

---

- [204] Noel M O'Boyle. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *Journal of Cheminformatics*, 4(1):22, 2012.
- [205] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. InChI - the worldwide chemical structure identifier standard. *Journal of cheminformatics*, 5(1):7, January 2013.
- [206] IUPAC. InChI FAQ.
- [207] Stefan Kuhn, Tobias Helmus, Robert J Lancashire, Peter Murray-Rust, Henry S Rzepa, Christoph Steinbeck, and Egon L Willighagen. Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML vocabulary for spectral data. *Journal of chemical information and modeling*, 47(6):2015–34, 2007.
- [208] Peter Ertl and Bernhard Rohde. The Molecule Cloud - compact visualization of large collections of molecules. *Journal of Cheminformatics*, 4(1):12, 2012.
- [209] Martin Gütlein, Andreas Karwath, and Stefan Kramer. CheS-Mapper - Chemical Space Mapping and Visualization in 3D. *Journal of cheminformatics*, 4(1):7, January 2012.
- [210] David Hoksza, Petr koda, Milan Vorilák, and Daniel Svozil. Molpher: a software framework for systematic chemical space exploration. *Journal of Cheminformatics*, 6(1):7, 2014.
- [211] Vincent Le Guilloux, Lionel Colliandre, Stéphane Bourg, Guillaume Guénégou, Julie Dubois-Chevalier, and Luc Morin-Allory. Visual characterization and diversity quantification of chemical libraries: 1. creation of delimited reference chemical subspaces. *Journal of chemical information and modeling*, 51(8):1762–74, August 2011.
- [212] Wendy A Warr and Wendy Warr. Integration, Analysis and Collaboration. An Update on Workflow and Pipelining in Cheminformatics. pages 1–7, 2007.
- [213] John Shon, Hitomi Ohkawa, and Juergen Hammer. Scientific workflows as productivity tools for drug discovery. *Current opinion in drug discovery & development*, 11(3):381–8, May 2008.
- [214] V. Curcin and M. Ghanem. Scientific workflow systems - can one size fit all? In *2008 Cairo International Biomedical Engineering Conference*, pages 1–9. IEEE, December 2008. ISBN 978-1-4244-2694-2.
- [215] M H Ebell. Visual programming languages. *M.D. computing : computers in medical practice*, 10(5):305–11, 1993.
- [216] Jeremy Goecks, Anton Nekrutenko, and James Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, January 2010.
- [217] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat, and Peter Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics (Oxford, England)*, 20(17):3045–54, November 2004.
- [218] Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007. ISBN 978-3-540-78239-1.
- [219] Michael A Kappler. Software for rapid prototyping in the pharmaceutical and biotechnology industries. *Current opinion in drug discovery & development*, 11(3):389–92, May 2008.
- [220] Andreas Hildebrandt, Anna Katharina Dehof, Alexander Rurainski, Andreas Bertsch, Marcel Schumann, Nora C Toussaint, Andreas Moll, Daniel Stöckel, Stefan Nickels, Sabine C Mueller, Hans-Peter Lenhof, and Oliver Kohlbacher. BALL—biochemical algorithms library 1.3. *BMC bioinformatics*, 11(1):531, January 2010.
- [221] Carole a Goble, Jiten Bhagat, Sergejs Aleksejevs, Don Cruickshank, Danus Michaelides, David Newman, Mark Borkum, Sean Bechhofer, Marco Roos, Peter Li, and David De Roure. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(Web Server issue):W677–82, July 2010.
- [222] Bernd Jagla, Bernd Wiswedel, and Jean-Yves Coppée. Extending KNIME for next-generation sequencing data analysis. *Bioinformatics (Oxford, England)*, 27(20):2907–9, October 2011.
- [223] Pierre Lindenbaum, Solena Le Scouarnec, Vincent Portero, and Richard Redon. Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME. *Bioinformatics (Oxford, England)*, 27(22):3200–1, November 2011.



## REFERENCES

---

- [224] Hendrik Strobel, Enrico Bertini, Joachim Braun, Oliver Deussen, Ulrich Groth, Thomas U Mayer, and Dorit Merhof. HiTSEE KNIME: a visualization tool for hit selection and analysis in high-throughput screening experiments for the KNIME platform. *BMC bioinformatics*, 13 Suppl 8(Suppl 8):S4, January 2012.
- [225] KNIME. Konstanz Information Miner - Professional Open-Source Software.
- [226] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10):e1003285, October 2013.
- [227] Mingshe Zhu, Haiying Zhang, and W Griffith Humphreys. Drug metabolite profiling and identification by high-resolution mass spectrometry. *The Journal of biological chemistry*, 286(29):25419–25, July 2011.
- [228] Arno Knorr, Aurelien Monge, Markus Stueber, André Stratmann, Daniel Arndt, Elyette Martin, and Pavel Pospisil. Computer-assisted structure identification (CASI)—an automated platform for high-throughput identification of small molecules by two-dimensional gas chromatography coupled to mass spectrometry. *Analytical chemistry*, 85(23):11216–24, December 2013.
- [229] Brian Goetz, Tim Peierls, Joshua Bloch, Joseph Bowbeer, David Holmes, and Doug Lea. *Java Concurrency in Practice*. Addison Wesley, 1 edition, 2006. ISBN 0321349601.
- [230] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics (Oxford, England)*, 24(21):2534–6, November 2008.
- [231] Ralf Tautenhahn, Christoph Böttcher, and Steffen Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, 9:504, January 2008.
- [232] Marcel van Herk. A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. *Pattern Recognition Letters*, 13(7):517–521, July 1992.
- [233] W Windig and W F Smith. Chemometric analysis of complex hyphenated data. Improvements of the component detection algorithm. *Journal of chromatography. A*, 1158(1-2):251–7, July 2007.
- [234] E Michael Thurman and Imma Ferrer. The isotopic mass defect: a tool for limiting molecular formulas by accurate mass. *Analytical and bioanalytical chemistry*, 397(7):2807–16, August 2010.
- [235] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie-Aude Coutouly, Ian Forsythe, Peter Tang, Savita Shrivastava, Kevin Jeroncic, Paul Stothard, Godwin Amegbey, David Block, David D Hau, James Wagner, Jessica Miniaci, Melisa Clements, Mulu Gebremedhin, Natalie Guo, Ying Zhang, Gavin E Duggan, Glen D Macinnis, Alim M Weljie, Reza Dowlatabadi, Fiona Bamforth, Derrick Clive, Russ Greiner, Liang Li, Tom Marrie, Brian D Sykes, Hans J Vogel, and Lori Querengesser. HMDB: the Human Metabolome Database. *Nucleic acids research*, 35(Database issue):D521–6, January 2007.
- [236] M Eichler and Till Francke. The GOLM-database standard—a framework for time-series data management based on free software. *EGU General Assembly ...*, 11:EGU2009–8070, 2009.
- [237] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry : JMS*, 45(7):703–14, July 2010.
- [238] Colin A Smith, Grace O’Maille, Elizabeth J Want, Chuan Qin, Sunia A Trauger, Theodore R Brandon, Darlene E Custodio, Ruben Abagyan, and Gary Siuzdak. METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring*, 27:747–751, 2005.
- [239] C J Van Rijsbergen. *Information Retrieval*, volume 30. 1979. ISBN 0408709294.
- [240] Tobias Kind and Oliver Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, 8:105, January 2007.
- [241] Tobias Kind and Oliver Fiehn. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics*, 7:234, January 2006.
- [242] Richard Baran and Trent R Northen. Robust automated mass spectra interpretation and chemical formula calculation using mixed integer linear programming. *Analytical chemistry*, 85(20):9777–84, October 2013.

## REFERENCES

---

- [243] Steffen Neumann and Sebastian Böcker. Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules. *Analytical and bioanalytical chemistry*, pages 2779–2788, October 2010.
- [244] Angelika Galezowska, Mark W Harrison, Julie M Herniman, Chris-Kriton Skylaris, and G John Langley. A predictive science approach to aid understanding of electrospray ionisation tandem mass spectrometric fragmentation pathways of small molecules using density functional calculations. *Rapid communications in mass spectrometry : RCM*, 27(9):964–970, May 2013.
- [245] Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, and Steffen Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, 11:148, January 2010.
- [246] Piotr T. Kasper, Miguel Rojas-Chertó, Robert Mistrik, Theo Reijmers, Thomas Hankemeier, and Rob J. Vreeken. Fragmentation trees for the structural characterisation of metabolites. *Rapid communications in mass spectrometry : RCM*, 26(19):2275–86, October 2012.
- [247] Gelio Alves, Aleksey Y Ogurtsov, and Yi-Kuo Yu. RAId\_aPS: MS/MS analysis with multiple scoring functions and spectrum-specific statistics. *PLoS one*, 5(11):e15438, January 2010.
- [248] Tomáš Pluskal, Taisuke Uehara, and Mitsuhiro Yanagida. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Analytical chemistry*, 84(10):4396–403, May 2012.
- [249] Harald Barsnes, Ingvar Eidhammer, and Lennart Martens. FragmentationAnalyzer: an open-source tool to analyze MS/MS fragmentation data. *Proteomics*, 10(5):1087–90, March 2010.
- [250] Sven Degroeve and Lennart Martens. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics (Oxford, England)*, 29(24):3199–203, December 2013.
- [251] Murray-Rust P and Rzepa HS. The Next Big Thing: From Hypermedia to Datuments. *Journal of Digital Information*, 5:248, 2004.
- [252] Barend Mons, Herman van Haagen, Christine Chichester, Peter-Bram 'T Hoen, Johan T den Dunnen, Gertjan van Ommen, Erik van Mulligen, Bharat Singh, Rob Hooft, Marco Roos, Joel Hammond, Bruce Kiesel, Belinda Giardine, Jan Velterop, Paul Groth, and Erik Schultes. The value of data. *Nature genetics*, 43(4):281–3, January 2011.
- [253] David Shotton, Katie Portwin, Graham Klyne, and Alistair Miles. Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS computational biology*, 5(4):e1000361, April 2009.
- [254] Henry S Rzepa. Chemical datuments as scientific enablers. *Journal of Cheminformatics*, 4(1):30, 2012.
- [255] Henry S Rzepa. The past, present and future of Scientific discourse. *Journal of cheminformatics*, 3(1):46, January 2011.
- [256] Bjorn J a Berendsen, Linda a M Stolker, and Michel W F Nielen. The (un)certainly of selectivity in liquid chromatography tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 24(1):154–63, January 2013.
- [257] Kerstin Scheubert, Franziska Hufsky, and Sebastian Böcker. Computational mass spectrometry for small molecules. *Journal of cheminformatics*, 5(1):12, March 2013.
- [258] Lars Ridder, Justin Johan Jozias van der Hooft, Stefan Verhoeven, Ric C H de Vos, Raoul J Bino, and Jacques Vervoort. Automatic chemical structure annotation of an LC-MS(n) based metabolic profile from green tea. *Analytical chemistry*, 85(12):6033–40, June 2013.
- [259] Thomas H Miller, Alessandro Musenga, David A Cowan, and Leon P Barron. Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks. *Analytical chemistry*, 85(21):10330–7, November 2013.
- [260] Haiwei Gu, G A Nagana Gowda, Fausto Carnevale Neto, Mark R Opp, and Daniel Raftery. RAMSY: ratio analysis of mass spectrometry to improve compound identification. *Analytical chemistry*, 85(22):10771–9, November 2013.
- [261] Tim De Meyer, Davy Sinnaeve, Bjorn Van Gasse, Elena Tsiporkova, Ernst R Rietzschel, Marc L De Buyzere, Thierry C Gillebert, Sofie Bekaert, José C Martins, and Wim Van Criekinge. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical chemistry*, 80(10):3783–90, May 2008.

## REFERENCES

---

- [262] Sophie Bourcier and Yannik Hoppilliard. Use of diagnostic neutral losses for structural information on unknown aromatic metabolites: an experimental and theoretical study. *Rapid Communications in Mass Spectrometry*, 23(1):93–103, 2009.
- [263] Markus Heinonen, Huibin Shen, Nicola Zamboni, and Juho Rousu. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics (Oxford, England)*, 28(18):2333–41, September 2012.
- [264] Stephen E. Stein and Donald R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, September 1994.
- [265] S E Stein. Estimating probabilities of correct identification from results of mass spectral library searches. *Journal of the American Society for Mass Spectrometry*, 5(4):316–23, April 1994.
- [266] Jaesik Jeong, Xue Shi, Xiang Zhang, Seongho Kim, and Changyu Shen. An empirical Bayes model using a competition score for metabolite identification in gas chromatography mass spectrometry. *BMC bioinformatics*, 12(1):392, October 2011.
- [267] Melissa a Bodnar Willard, Ruth Waddell Smith, and Victoria L McGuffin. Statistical approach to establish equivalence of unabbreviated mass spectra. *Rapid communications in mass spectrometry : RCM*, 28(1): 83–95, January 2014.
- [268] Fumio Matsuda, Hiroshi Tsugawa, and Eiichiro Fukusaki. Method for assessing the statistical significance of mass spectral similarities using basic local alignment search tool statistics. *Analytical chemistry*, 85(17):8291–7, September 2013.
- [269] Kenneth Manning, Mahmut Tör, Mervin Poole, Yiguo Hong, Andrew J Thompson, Graham J King, James J Giovannoni, and Graham B Seymour. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nature genetics*, 38(8):948–52, August 2006.
- [270] S. Moore. Use of genomics tools to isolate key ripening genes and analyse fruit maturation in tomato. *Journal of Experimental Botany*, 53(377):2023–2030, October 2002.
- [271] Julia Vrebalov, Diane Ruezinsky, Veeraragavan Padmanabhan, Ruth White, Diana Medrano, Rachel Drake, Wolfgang Schuch, and Jim Giovannoni. A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (rin) locus. *Science (New York, N.Y.)*, 296(5566):343–6, April 2002.
- [272] Philippe Rocca-Serra, Marco Brandizi, Eamonn Maguire, Nataliya Sklyar, Chris Taylor, Kimberly Begley, Dawn Field, Stephen Harris, Winston Hide, Oliver Hofmann, Steffen Neumann, Peter Sterk, Weida Tong, and Susanna-Assunta Sansone. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics (Oxford, England)*, 26:2354–2356, 2010.
- [273] Johannes Griss, Timo Sachsenberg, Mathias Walzer, Oliver Kohlbacher, Andrew R. Jones, Henning Hermjakob, and Juan Antonio Vizcaino. mzTab: exchange format for proteomics and metabolomics results. Technical report, PSI Proteomics Informatics Workgroup, 2013.
- [274] Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, and Christoph Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(Database issue):D456–63, January 2013.
- [275] Francesc Fernández-Albert, Rafael Llorach, Cristina Andres-Lacueva, and Alexandre Perera-Lluna. Peak Aggregation as an Innovative Strategy for Improving the Predictive Power of LC-MS Metabolomic Profiles. *Analytical chemistry*, February 2014.
- [276] V Pellegrin. Molecular formulas of organic compounds: the nitrogen rule and degree of unsaturation. *Journal of Chemical Education*, pages 626–633, 1983.
- [277] C. Benecke, T. Grüner, A. Kerber, R. Laue, and T. Wieland. MOLEcular structure GENeration with MOLGEN, new features and future developments. *Fresenius' Journal of Analytical Chemistry*, 359(1): 23–32, August 1997.
- [278] Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711, January 2008.
- [279] W. J. Krzanowski. Cross-Validation in Principal Component Analysis. *Biometrics*, 43(3):575, September 1987.

## REFERENCES

---

- [280] Anatoly P. Sobolev, Donatella Capitani, Donato Giannino, Chiara Nicolodi, Giulio Testone, Flavio Santoro, Giovanna Frugis, Maria a. Iannelli, Autar K. Mattoo, Elvino Brosio, Raffaella Gianferri, Irene DAMico, and Luisa Mannina. NMR-Metabolic Methodology in the Study of GM Foods. *Nutrients*, 2(1): 1–15, January 2010.
- [281] Gemma Oms-Oliu, M.L.a.T.M. Hertog, B. Van de Poel, J. Ampofo-Asiama, A.H. Geeraerd, and B.M. Nicolai. Metabolic characterization of tomato fruit during preharvest development, ripening, and postharvest shelf-life. *Postharvest Biology and Technology*, 62(1):7–16, October 2011.
- [282] Autar K Mattoo, Anatoli P Sobolev, Anil Neelam, Ravinder K Goyal, Avtar K Handa, and Anna L Segre. Nuclear magnetic resonance spectroscopy-based metabolite profiling of transgenic tomato fruit engineered to accumulate spermidine and spermine reveals enhanced anabolic and nitrogen-carbon interactions. *Plant physiology*, 142(4):1759–70, December 2006.
- [283] Sonia Osorio, Rob Alba, Cynthia M B Damasceno, Gloria Lopez-Casado, Marc Lohse, Maria Inés Zanor, Takayuki Tohge, Björn Usadel, Jocelyn K C Rose, Zhangjun Fei, James J Giovannoni, and Alisdair R Fernie. Systems biology of tomato fruit development: combined transcript, protein, and metabolite analysis of tomato transcription factor (nor, rin) and ethylene receptor (Nr) mutants reveals novel regulatory interactions. *Plant physiology*, 157(1):405–25, September 2011.
- [284] Xinhua Zhang, Lin Shen, Fujun Li, Demei Meng, and Jiping Sheng. Hot air treatment-induced arginine catabolism is associated with elevated polyamines and proline levels and alleviates chilling injury in postharvest tomato fruit. *Journal of the science of food and agriculture*, 93(13):3245–51, October 2013.
- [285] Emma L Schymanski and Steffen Neumann. CASMI: And the Winner is . . . *Metabolites*, 3(2):412–39, January 2013.
- [286] Emma L Schymanski and Steffen Neumann. The Critical Assessment of Small Molecule Identification (CASMI): Challenges and Solutions. *Metabolites*, 3(3):517–38, January 2013.
- [287] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of chemical information and computer sciences*, 43(2):493–500, 2003.
- [288] Vincent Le Guilloux, Lionel Colliandre, Stéphane Bourg, Guillaume Guénégou, Julie Dubois-Chevalier, and Luc Morin-Allory. Visual characterization and diversity quantification of chemical libraries: 1. creation of delimited reference chemical subspaces. *Journal of chemical information and modeling*, 51(8):1762–74, August 2011.
- [289] Wendy a Warr. Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of computer-aided molecular design*, 26(7):801–4, July 2012.
- [290] Stefan Krause, Egon Willighagen, and Christoph Steinbeck. JChemPaint - Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures. *Molecules*, 5(1):93–98, January 2000.
- [291] The Jmol Team. Jmol: an open-source Java viewer for chemical structures in 3D., 2007.
- [292] Nina Jeliaskova and Vedrin Jeliaskov. AMBIT RESTful web services: an implementation of the OpenTox application programming interface. *Journal of Cheminformatics*, 3(1):18, 2011.
- [293] Nikolay T. Kochev, Vesselina H. Paskaleva, and Nina Jeliaskova. Ambit-Tautomer: An Open Source Tool for Tautomer Generation. *Molecular Informatics*, 32(5-6):481–504, June 2013.
- [294] Paula de Matos, Rafael Alcántara, Adriano Dekker, Marcus Ennis, Janna Hastings, Kenneth Haug, Inmaculada Spiteri, Steve Turner, and Christoph Steinbeck. Chemical Entities of Biological Interest: an update. *Nucleic acids research*, 38(Database issue):D249–54, January 2010.
- [295] Renxiao Wang, Ying Gao, and Luhua Lai. Calculating partition coefficient by atom-additive method. *Persp. Drug Discov. Design*, 19:47–66, 2000.
- [296] Christopher A. Lipinski. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, December 2004.
- [297] P Wexler. TOXNET: an evolving web resource for toxicology and environmental health information. *Toxicology*, 157(1-2):3–10, January 2001.
- [298] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of chemical information and modeling*, June 2012.

## REFERENCES

---

- [299] Nan An, Farid Van Der Mei, and Adelina Voutchkova-Kostal. Global Model for Octanol-Water Partition Coefficients from Proton Nuclear Magnetic Resonance Spectra. *Molecular Informatics*, 33(4):286–292, April 2014.
- [300] Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45(12):2615–23, June 2002.
- [301] Indigo Toolkit: a universal organic chemistry toolkit. GGA Software Services.