# Method Development in Metabolomics

**Dissertation**

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Sc. Oliver Alka

aus Schweinfurt

Tübingen

2023

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

| | |
|---|---|
| Tag der mündlichen Qualifikation: | 13.12.2023 |
| Dekan: | Prof. Dr. Thilo Stehle |
| 1. Berichterstatter/-in: | Prof. Dr. Oliver Kohlbacher |
| 2. Berichterstatter/-in: | Prof. Dr. Sven Nahnsen |

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

*Method Development in Metabolomics*

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Ort, Datum

Unterschrift

"*Act so that the effects of your action are compatible with the permanence of genuine human life*"

Hans Jonas (1903-1993)

# Abstract

The field of metabolomics is concerned with analyzing data from high-throughput experiments. Its objective is the identification, quantification, and elucidation of the function and interaction of small molecules in a biological system. The prevalent methods used in metabolomics are nuclear magnetic resonance spectroscopy and mass spectrometry. A typical mass spectrometry metabolomics analysis workflow is composed of several steps. First, biological samples are measured using liquid chromatography and mass spectrometry. Second, computational mass spectrometry is used to analyze the acquired data. The results are then stored in a human-readable format, statistically post-processed, and visualized. The driving force of the field is the development of new methods on the analytical and computational side to reach the above-mentioned aims. Nonetheless, there are still some major unsolved issues at different stages of the analysis workflow.

Controlling the false-discovery rate (FDR) is well established in other fields (i.e., proteomics), but so far, methods are lacking in the field of metabolomics. This seriously limits the confidence in reported identifications and quantifications, and manual assessment is still common practice. In recent years different methods have been established for untargeted approaches [1,2]. However, in terms of targeted strategies, the lack of robust FDR estimators prevented the field from obtaining highly confident quantifications. Progress in automating the manual process is substantial to advance targeted metabolomics research and allow proper high-throughput analysis. We established an automated, FDR-controlled targeted analysis workflow that enables a robust FDR estimation for the first time, thus improving the comparability of results in the metabolomics field.

Another critical aspect of scientific research is representing and sharing analysis results based on the FAIR principles. The FAIR principles stand for findable, accessible, interoperable, and reusable [3,4]. In 2014, the human-readable file format *MzTab* was introduced in the proteomics and metabolomics fields to enable the distribution of analysis results in a standardized open format [5]. However, in recent years, the limitations of this format regarding metabolomics data have become apparent [6]. As part of the Proteomics Standard Initiative, we designed the improved standard *MzTab-M* that focuses on interoperability and reusability and integrated it into our OpenMS software framework.

Metabolomics has a massive range of applications and can be used to answer a variety of scientific questions. The field attempts to answer individual data- and objective-related issues by developing new problem-specific post-processing methods, as we show based on an example in the area of food chemistry. In recent years, the production of primary cacao products, such as cacao butter, moved from Europe to the cacao-producing countries. This leads to the challenge of shifting the quality assessment from raw to primary products to uphold the quality standards and control in the European market. To this end, we provided the basis for such a method by using biomarker identification and machine learning. Using a regression method, we were able to assess the shell quantity in a mixture of bean and shell and, with it, the quality of the product.

# Zusammenfassung

Der Forschungszweig der Metabolomik beschäftigt sich mit der Analyse von Stoffwechselprodukten in biologischen Systemen, sogenannten Metaboliten. Die zu analysierenden Daten werden oft mit Hilfe von massenspektrometrischen Hochdurchsatzmethoden generiert. Die Identifizierung und Quantifizierung der Metabolite im biologischen System stehen hierbei im Vordergrund. Des Weiteren trägt die Methodik zur Funktions- und Interaktionsaufklärung bei. Der Ablauf eines solchen Experiments kann in mehreren grundlegenden Schritten zusammengefasst werden. Zunächst wird eine biologische Probe mit Hilfe von Flüssigchromatographie gekoppelt an Massenspektrometrie analysiert. Danach folgt die computergestützte Analyse der gemessenen Daten. Anschließend werden die Daten in einem menschenlesbaren Format gespeichert, welches für weitere statistische Analysen und Visualisierungen genutzt werden kann. Die treibende Kraft des Forschungsfeldes ist die Weiter- und Neuentwicklung experimenteller und computergestützter Analysemethoden, um die oben genannten Ziele zu erreichen. Einige Stellen des Arbeitsablaufs sind jedoch noch problembehaftet.

Die Schätzung und Anwendung einer *False Discovery Rate* (FDR) ist in der Proteomik und anderen Forschungsfeldern weit verbreitet. Diese hilft bei der Abschätzung der Zahl an möglichen falsch positiven Ergebnissen im gesamten Ergebnisraum und deren Kontrolle. Jedoch finden diese Methoden in der Metabolomik derzeit kaum Verwendung. Dies mindert das Vertrauen in die Ergebnisse der Identifizierung und Quantifizierung und folglich sind manuelle Validierungen gängige Praxis. In den letzten Jahren wurden erfolgreich einige Methoden für die ungezielte metabolomische Analyse entwickelt [1,2], welche auf die Identifizierung und Quantifizierung möglichst vieler bekannter und unbekannter Metabolite in einem Experiment abzielt. Im gezielten Einsatzbereich, welcher genutzt wird, um bereits bekannte Metabolite zu quantifizieren, fehlt die robuste Schätzung der FDR gänzlich. Für den weiteren Fortschritt ist es daher unumgänglich, die derzeitigen manuellen Prozesse zu automatisieren und eine FDR bereitzustellen. Dieser Fragestellung haben wir uns angenommen und präsentieren einen automatisierten Workflow für die Analyse von gezielten Experimenten mit der Möglichkeit zur FDR-Schätzung und Kontrolle. Dieser erlaubt die Abschätzung einer robusten FDR und fördert somit die Vergleichbarkeit zwischen verschiedenen Experimenten.

Ein weiterer kritischer Aspekt im Forschungsfeld ist die Repräsentation der Ergebnisse, basie-

rend auf den *FAIR* Prinzipien. Diese beschreiben allgemeingültige Prinzipien in der Forschung von Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit[3,4]. Um die Verteilung und Wiederverwendbarkeit von Ergebnissen in den Forschungsfelder der Metabolomik und Proteomik zu gewährleisten, wurde im Jahr 2014 das öffentliche, standardisierte menschenlesbare Format *MzTab* eingeführt[5]. In den letzten Jahren kristalisierte sich heraus, dass dieses Format für den Einsatz in der Metabolomik nicht vollumfänglich geeignet ist[6]. Um die Interoperabilität und die Wiederverwendbarkeit zu garantieren, hat sich die *Proteomics Standard Initiative* zu einer Weiterentwicklung des Formats entschieden. Wir haben beim Entwurf des neuen Formats *MzTab-M* mitgewirkt und dieses in unser Software System *OpenMS* integriert.

Mit Hilfe der Metabolomik lassen sich Fragestellungen aus verschiedensten Bereichen beantworten, wie wir an einem Beispiel aus der Lebensmittelindustrie zeigen. In den letzten Jahren hat sich die Produktion primärer Kakaoprodukte, beispielsweise der Kakaobutter, aus Europa direkt in die entsprechenden Anbaugebiete verlagert. Um dem Problem fallender Qualitätsstandards und dem Wechsel von der Qualitätsprüfung des Roh- zum Primärprodukt zu begegnen, müssen neue Methoden entwickelt werden um die Qualitätskontrollen in den europäischen Ländern zu ermöglichen. Wir haben Biomarker-Identifizierung und maschinelles Lernen genutzt, um eine solche Methode zu etablieren. Diese kann durch die Nutzung einer linearen Regression den Anteil von Schalenbestandteilen in einer Mischung aus Bohne und Schale und somit die Qualität des Produkts bestimmen.

# Acknowledgments

I would like to thank my advisor, Prof. Oliver Kohlbacher, for the opportunity to achieve my Ph.D. in his group. I am grateful to him and to Prof. Hannes Röst for their guidance on my projects and the time that they spent discussing various drawbacks that I encountered, as well as troubleshooting.

I also want to thank my colleagues, especially Timo Sachsenberg and Julianus Pfeuffer, for their valuable help and the discussions held around software engineering, programming, and OpenMS. Fabian Aicheler, Leon Bichmann, Leon Kuchenbecker for the engaging conversations in the office, as well as my other colleagues for a productive and pleasant time. Special thanks go to my writing buddies Thorsten Tiede, Mirjam Figaschewski and Eftychia Kontou for their valuable input to this thesis.

In addition, I want to thank all my lab mates from Toronto, especially Premy Shanthamoorthy, Annie Ha, and Shubham Gupta, for the excellent collaboration and time spent together.

I am also grateful to Michael Witting and Karin Kleingrewe from Munich for the data acquisition and project discussions. As well as the Group of Prof. Sebastian Böcker in Jena, especially Markus Fleischauer und Marcus Ludwig for the excellent collaboration between OpenMS and SIRIUS.

I would like to express my appreciation to Tjeerd Dijkstra and George Rosenberger for their statistical advice throughout my Ph.D.

Special thanks to Nicolas Cain and Mark Rurik for their cooperation and help with the Foodomics projects.

Finally, I would like to express my gratitude to my parents, my brother, his family, my wife Lisa, and my Son Luke, who was born during my Ph.D. studies, for their motivation and support.

# Contents

# Chapter 1

# Introduction

## 1.1 Motivation

Method development is a foundation pillar of modern science. Pushing the boundaries of science with novel and improved methods paves the way for new insights and a better understanding of biological systems. One field, which vastly improved thanks to experimental and computational developments, is molecular biology. With its modern high-throughput -omics technologies, it concentrates on answering aspects of the central dogma of molecular biology, the flow of information in biological systems, and the understanding of biological machinery. The omics-associated fields aim to explain the biological association between DNA (genomics), RNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). In short, DNA transcribes into RNA, which translates into proteins. Many proteins have enzymatic/catalytic functions facilitating metabolic processes.
Compared to other -omics fields, metabolomics is relatively young. Its ultimate objective is to identify and quantify the metabolome, the complete set of metabolites in an organism, a specific tissue, or a cell at a specific point in time. These metabolites represent the biochemical activity of an organism and describe its molecular phenotype. Over the last decades, the methods in metabolomics advanced from *paper chromatography* to *gas chromatography-coupled mass spectrometry* (GC-MS), *nuclear magnetic resonance spectroscopy* and *liquid chromatography-coupled mass spectrometry* (LC-MS). New analytical and computational methods were developed alongside these advances in technology, which enables the high-throughput analysis of metabolomics data.
Focusing on LC-MS, two main strategies are commonly applied depending on the experimental context. One is a targeted strategy, the so-called *metabolic profiling*, that tries to quantify known analytes over a wide dynamic range. The other is an untargeted strategy called *metabolic fingerprinting* that tries to identify and quantify unknown metabolites (such as novel biomarkers) by comparing conditions. Unfortunately, automation of correct metabolite identification is still

an unresolved problem in the field. Although organic compounds are usually composed of the chemical elements C, H, O, N, P, and S, the lack of common building blocks - in contrast to peptides - leads to an enormous structural diversity of chemical compounds. This diversity makes it hard to identify analytes or predict their fragmentation patterns. In addition, the association of varying adducts to the compounds during the experimental process further complicates the identification.
Controlling the FDR is well established in other fields (e.g., proteomics), but so far, such methods are lacking in the field of metabolomics. This seriously limits the confidence in reported identifications and makes manual assessment still a common practice.
In recent years, different methods have been established for untargeted approaches[1,2]. However, in terms of targeted strategies, the lack of robust FDR estimators prevented the field from obtaining highly confident quantifications. It is thus essential to make progress in automating the manual process in order to advance the targeted field and allow proper high-throughput analysis.
Another critical aspect of scientific research is representing and sharing analysis results based on the FAIR principles. The FAIR principles stand for findable, accessible, interoperable, and reusable[3]. In 2014, the Proteomics Standards Initiative introduced the distribution of proteomics and metabolomics analysis results in a standardized open format called *MzTab*[5]. However, in recent years, the limitations of the format regarding metabolomics data have become apparent[6]. Therefore, there is a need for the development of an improved standard focusing on interoperability and reusability.
The field of metabolomics has a wide range of applications and can be used to answer a variety of scientific questions. A well-known field of application is food chemistry. For example, in recent years, the production of primary cacao products, such as cacao butter, migrated from Europe to the cacao-producing countries. This relocation led to the issue of upholding the quality standards of cacao products in Europe. To this end, a new method needs to be established to allow for quality assessment of such products.

## 1.2 Structure of this Thesis

In this thesis, we present three methods developed in different stages of the metabolomics analysis workflow (Fig. 1.1). First, in a typical workflow, an experiment is performed and measured using LC-MS-MS/MS. Second, the data is analyzed using computational mass spectrometry methods. Finally, reported identification and quantification results are further statistically post-processed and visualized. Following the introduction, Chapter 2 establishes the relevant technical and biological background for the metabolomics analysis workflow.
Afterwards, we present a new way of computational mass spectrometry analysis by combining the data from untargeted metabolite fingerprinting and targeted metabolite profiling experi-

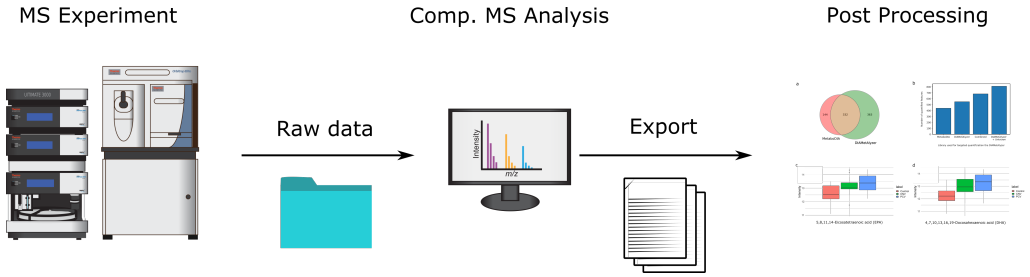


**Figure 1.1: Metabolomics Analysis Workflow** Biological samples are measured using LC-MS-MS/MS. Computational mass spectrometry methods analyse the acquired data from the instrument. Further post-processing of the exported results can include statistical evaluation and generation of visual representations.

ments. We introduce a novel automated workflow for the targeted analysis with FDR control in metabolomics in Chapter 3. Here, metabolite quantification results are subjected to robust FDR control. Our approach leads to quantitative matrices with fewer missing values while controlling the FDR, which improves biologically relevant findings compared to other methods. In addition, we contribute to the standard specification and its documentation, as well as implement an advanced reporting format for metabolomics data called *MzTab-M* into our OpenMS software library[7] (Chapter 4). This greatly enhances interoperability and reusability in the scientific community. *MzTab-M* is an improved standardised format based on the previous standard *MzTab*.

In Chapter 5, we focus on food fingerprinting and biomarker discovery on the example of improved quality assessment of the primary product of cacao production. We show that the application of our methods enables statistically valid and reproducible results.

Chapter 6 concludes this thesis.

# Chapter 2

# Background

## 2.1 Metabolomics

Metabolomics is the field of research that studies the quantity, function, and interaction of small molecules in a biological system during the alternation with a stimulus[8,9]. The *Metabolome* describes the totality of small molecules found in an organism, cell, or tissue under specified conditions[10]. Due to the relationship between the biological activity of a metabolite and the biological phenotype, metabolomics is used, for example, in personalized medicine, biomarker discovery, the study of natural products, and food profiling[11,12,13,14,15]. Mass spectrometry (MS) combined with chromatographic analyte separation is the primary analytical technique for large scale, high-throughput experiments. There are two analysis strategies in the field, targeted and untargeted metabolomics. The first one aims at detecting and quantifying a list of known, annotated, and characterized metabolites focused on a specific research question[16], whereas the second aims to detect, identify and quantify as many metabolites as possible in a sample, including chemically unknown compounds[17].

## 2.2 Mass Spectrometry-based Metabolomics

The scientific question defines the experimental design of any MS experiment, nevertheless all experiments share the same fundamental setup. In a sample preparation step, analytes are extracted and prepared for MS. In most cases, separation techniques follow to reduce the complexity of the sample. Afterwards, a mass spectrometer measures the obtained analytes mass-to-charge ratio (m/z). Lastly, analytes are identified and quantified from the acquired raw data using computational methods.

### 2.2.1 Sample Preparation

Due to the chemical diversity of small molecules and their physicochemical properties, various methods exist to extract a certain population of metabolites from a sample. The choice of extraction method is dependent on the organism of interest and has a high impact on metabolome coverage and metabolite classes covered. In metabolomics, liquid-liquid extraction is found in most protocols[18]. A protocol for the extraction of medium-polarity metabolites, such as organic acids and amino acids, uses a mixture of ethanol and water for the initial extraction, followed by treatment with chloroform to remove low polarity interferences[19]. Ethanol- and methanol-based extractions lead to the best metabolite coverage regarding animal and plant tissue[18,20]. For further details and additional methods, the following reviews might be of interest[21,22,23].

### 2.2.2 Separation Techniques

Separation techniques are experimental approaches that allow for complexity reduction of a sample by separating mixtures into individual analytes. In combination with mass spectrometry, chromatography is one of the most widely used separation methods.

#### Chromatography

Chromatography separates a mixture into individual analytes by exploiting their different physicochemical properties. There are two well-known chromatographic methods used in metabolomics. The first one is gas chromatography (GC), which allows the analysis of volatile substances. The second one is liquid chromatography (LC), which enables the study of non-volatile substances, described in detail in the following section.

#### Liquid Chromatography

An LC system is comprised of a solvent, an eluent, a pump, an injection valve, a chromatographic column, and a detector (Fig. 2.1). A pump pumps a solvent (mobile phase) through a temperature-controlled column with chromatographic packing material (stationary phase). Then, an analyte mixture (i.e., biological sample) is injected using the injection valve. Next, the introduced analytes are separated by their physicochemical properties by an eluent concentration gradient and detected via a detector (e.g., ultraviolet-visible light spectroscopy). The columns stationary phase has specific physicochemical properties and is combined with a solvent that represents the mobile phase. The method is versatile due to the various stationary and mobile phase combinations. Normal-phase LC uses a polar stationary bed (e.g., silica) and a nonpolar liquid mobile phase (e.g., hexane). Here, analytes with high polarity are retained longer on the polar surface and elute at a later retention time than less polar analytes.

In this thesis, a technique, which uses opposite polarities, is mainly used. In reversed-phase chromatography, hydrophobic material, as a stationary phase (e.g., C18), separates analytes by hydrophobicity. In combination with a hydrophilic mobile phase (i.e., acetonitrile/water mixture), hydrophobic analytes show a higher affinity with the column and are therefore retained longer. A combination of pressure and columns with a smaller diameter, as well as pore size, are characteristics that enhance this setup, referred to as high-performance liquid chromatography (HPLC). HPLC increases the sensitivity by increasing the signal-to-noise ratio. The resolution of the chromatography is dependent on the column length and inner diameter, particle size, stationary phase, flow rate, pressure and temperature, and the composition of the mobile phase. In addition, optimization of the analytes retention time and resolution is possible by changing the length of the run (time) and the gradient method.

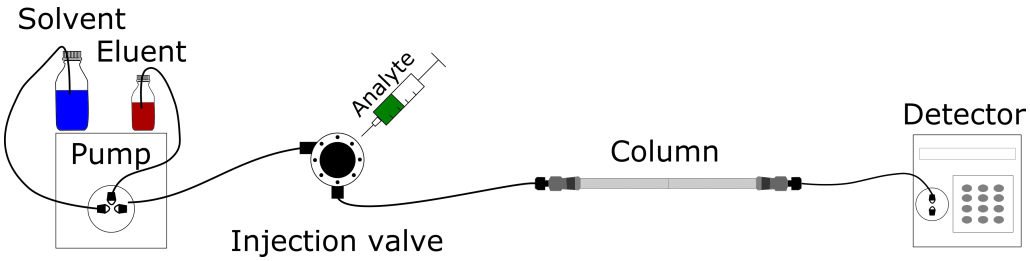


**Figure 2.1: High-Performance Liquid Chromatography** A high-pressure pump system pumps a solvent (mobile phase) through a column with chromatographic packing material (stationary phase). Then, an analyte mixture (i.e., biological sample) is injected using the injection valve. Next, the introduced analytes are separated by their physicochemical properties by an eluent concentration gradient, and detected via a detector (e.g., ultraviolet-visible light spectroscopy).

### 2.2.3 Mass Spectrometry

An MS measures the m/z ratio of ionized molecules. It consists of an ion source, a mass analyzer, and an ion detector. Figure 2.2 shows the schematics of a Quadrupole Time-of-Flight (qTOF) instrument. It consists of an electrospray ionization (ESI) ion source, three quadrupoles (Q1-3) mass analyzers coupled to a Time-of-Flight (TOF) detector.

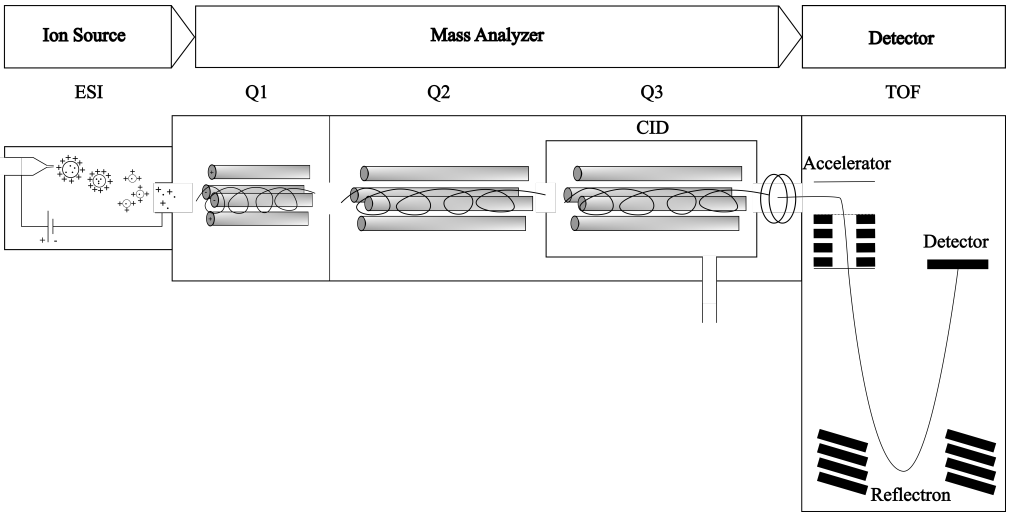


**Figure 2.2: Setup of a qTOF Mass Spectrometer** A mass spectrometer consists of three main components. First, the ion source ionizes the analytes. Second, the mass analyzer filters and sorts these analytes based on their m/z ratio. These ionized particles are then counted and recorded by the detector. The qTOF mass spectrometer consists of an electrospray ionization (ESI) ion source, three quadrupoles mass analyzers coupled to a Time-of-Flight (TOF) detector. The TOF part of the MS consists of an accelerator, a reflectron, and a detector.

### Ion Source

During the ESI process, developed in 1984, the analyte solution is sprayed into a vacuum through a charged capillary[24]. Due to the electric field, an excess of similarly charged particles is collected at the capillary tip. Their repulsion leads to the formation of a Taylor cone which means that the ionized analytes exit as charged aerosol droplets. Furthermore, evaporation leads to the shrinking of those droplets. Neutral gas can be used to support the evaporation process. A Coulomb explosion leads to even smaller droplets. This happens when the droplet radius hits the Rayleigh Limit[25]. It states that a droplet can only carry a specific amount of equal charge; when the radius falls below this threshold, it leads to its explosion. One of the theories leading to single ionized particles, the so-called charge residue model, describes an iterative cycle of evaporation and fission until on average one analyte remains in a droplet[26]. At the end of the process, the ionized particles enter the high vacuum parts of the MS via the continuous gas stream.

### Mass Analyzer

The qTOF instrument shown in Figure 2.2 has three quadrupole mass analyzers. These consist of four parallel metal rods, where an oscillating electric field is applied. The mass analyzer stabilizes the flight path of the ionized particles of a certain m/z ratio range. Due to their unstable trajectories, ions with m/z values outside the defined range are expelled unless the

frequency is changed, in order to allow them through the quadrupole. A quadrupole can also act as a detector, but in the qTOF instrument, it is used to forward the filtered ions to the TOF detector.

#### Detector

One of the most popular detectors is the TOF detector. Figure 2.2 shows an orthogonal acceleration TOF. The ions accelerate from an orthogonal entry point into a field-free drift zone. They hit a reflectron, which uses a constant electrostatic field to reflect them onto the detector. It records the time of flight between the extraction pulse and the hit onto the detector. The instrument calculates the m/z values from the flight duration and the applied acceleration voltage.

#### Mass Spectrometry Data

The MS measures a sample over the retention time of the chromatography and the scan m/z (i.e., 100 to 700) set on the instrument. The measurement of one sample is one MS run and produces an MS map (Fig. 2.3a). The map consists of individual mass spectra measured at a specific time. All ions measured in m/z and intensity for a given retention time are represented in a spectrum. As an example, the spectrum at retention time 461.92 s in the mass spectrometry map is visualized in Figure 2.3b.

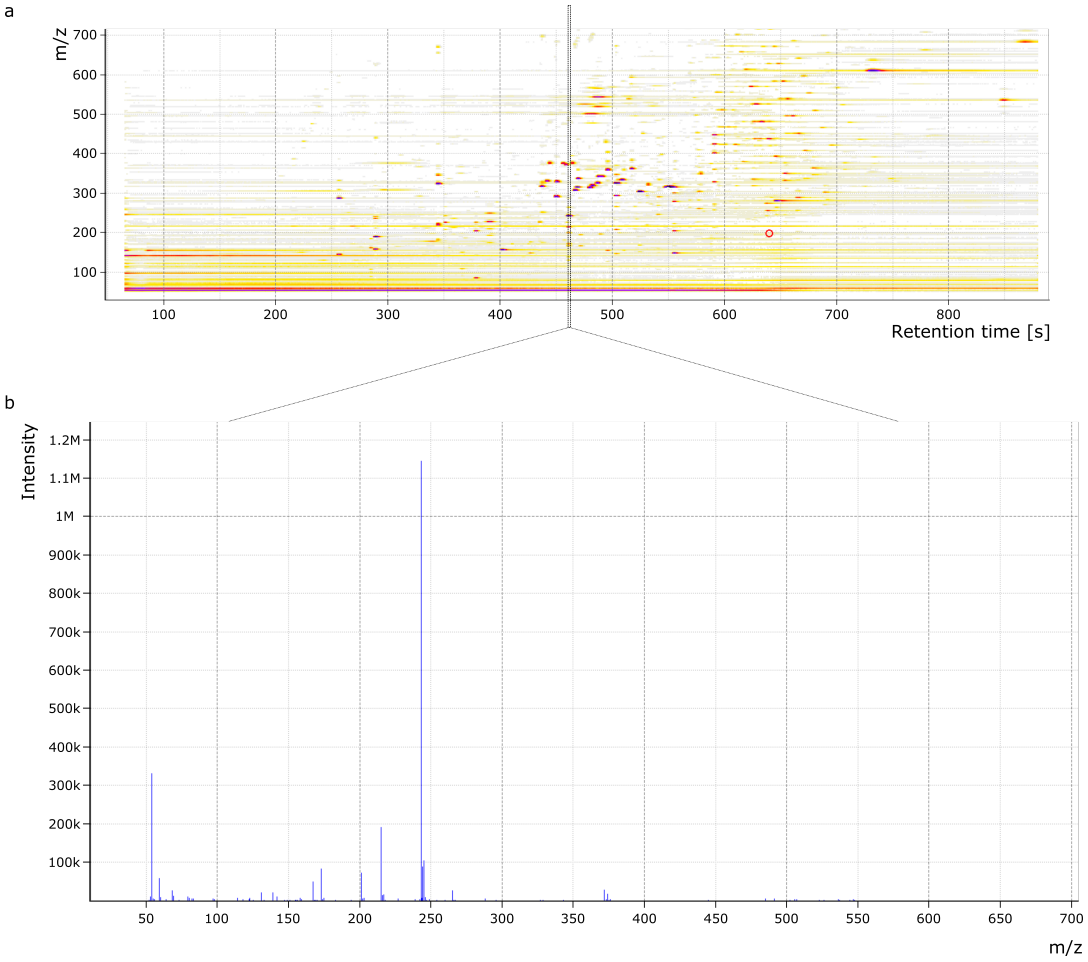


**Figure 2.3: Mass Spectrometry Data** a) The MS map shows the ions measured by m/z (Y-Axis) over the retention time (X-Axis), coloured by the signal intensity. b) The MS spectrum shows the ions measured by intensity (Y-Axis) over the m/z (X-Axis) at 461.92 seconds

### 2.2.4 Tandem Mass Spectrometry

In metabolomics, identification based on the MS1 m/z (accurate mass) and the isotope pattern, which occurs due to the natural abundance of isotopes (i.e., $^{13}C$), is often ambiguous. Tandem mass spectrometry, MS/MS, can be applied to assess the analytes substructure and improve its identification. To this end, a so-called precursor ion within a specific mass window (i.e., 0.7 m/z) is isolated and kinetically fragmented using an inert gas (i.e., Argon). Fragments produced by collision-induced dissociation (CID) are stored in an MS/MS (MS2) spectrum and provide information that helps to resolve the ambiguities in identification (Fig. 2.4).

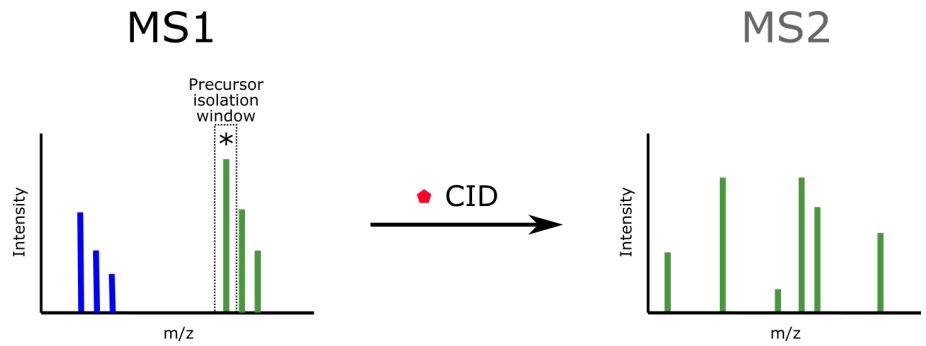


**Figure 2.4: Tandem Mass Spectrometry** Collision induced dissociation performed on a monoisotopic precursor ion in a specific predefined isolation window (i.e., 0.7 m/z) results in an MS2 spectrum. The MS/MS spectrum depicts fragment peaks (substructures) originating from the isolated analyte/precursor.

### Acquisition Methods

Three data acquisition methods are described in this section. Data-dependent acquisition (DDA), data-independent acquisition (DIA), and Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH). Figure 2.5 depicts a comparison of DDA and SWATH, which is an advanced DIA technique.

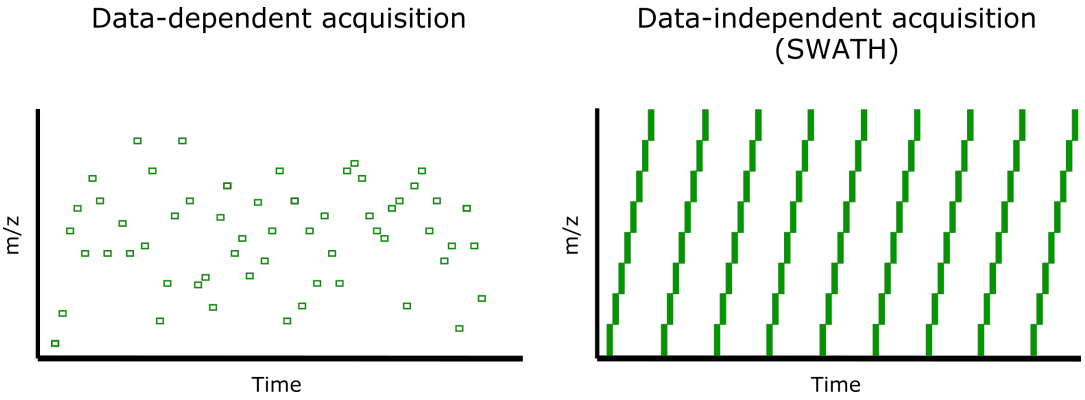


**Figure 2.5: Acquisition Methods** Semi-stochastic DDA: The instrument selects the $n$ highest intensity precursors iteratively for fragmentation. SWATH: Sequential iteration over smaller mass windows, i.e., eight windows with 125 Da in approximately one second. Allowing for fragmentation of all precursors in such a window.

### Data-Dependent Acquisition

DDA, called "shotgun" acquisition, is a semi-stochastic method, which selects the $n$ highest intensity precursors in an MS1 survey scan for MS2 acquisition (Fig. 2.6). This method references the precursor and its fragments with the drawback of introducing a bias to high-intensity peaks. The instrument decides for each sample which precursors are fragmented. Because of the stochastic nature, an analytes MS2 spectrum might not be recorded in every run, even if

the precursor is detected (undersampling). For that reason, DDA is regarded to be less precise and reproducible in contrast to DIA[27,28].

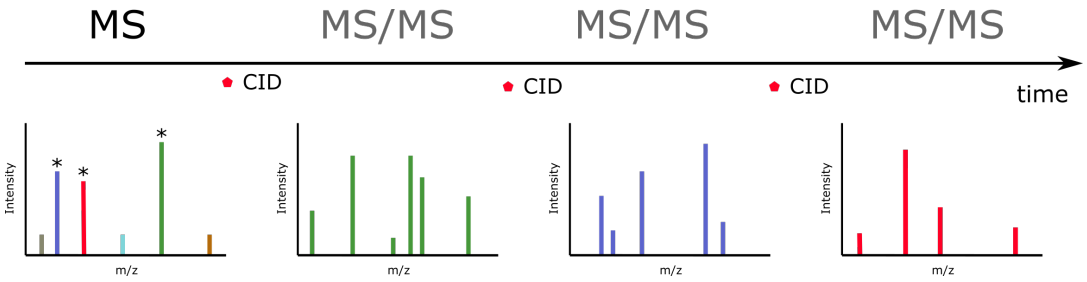


**Figure 2.6: Data-Dependent Acquisition** The top $n$ highest intensity precursors are sequentially fragmented and measured (MS2 spectrum). Thereby the reference between the precursor and its fragments is retained.

### Data-Independent Acquisition

DIA reduces this bias to high-intensity precursors by fragmenting all precursors in a given mass window. During a standard DIA procedure, all precursors in a given scan m/z range (i.e., 100 - 1200), detected in a MS1 survey scan are fragmented simultaneously. However, this increases the complexity of the MS2 spectra drastically, and the reference between precursor and fragments is lost (Fig 2.7). In gerneral, DIA allows the quantification of complex mixtures of analytes over an extensive dynamic range. Furthermore, in contrast to DDA, DIA overcomes the challenge of undersampling[27,28]. Another advantage is that the data can be reanalyzed with improved algorithms since the whole MS map is covered during the analysis.

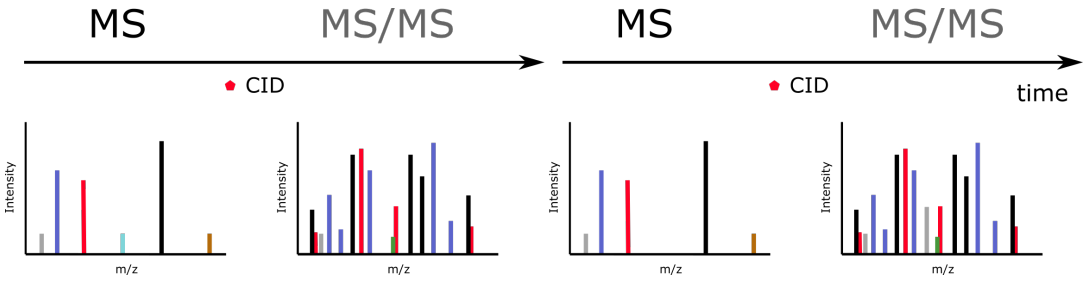


**Figure 2.7: Data-Independent Acquisition** All precursors belonging into a user-defined m/z range in the MS1 spectrum are fragmented simultaneously, leading to a complex MS2 spectrum. Thus, the challenge of DDA undersampling is solved, with the drawback of losing the precursor-fragments reference.

One of the most popular DIA methods is SWATH, which reduces the complexity of mulitplexed spectra, in contrast to the standard method, to allow for a more accessible analysis. This is achieved by reducing the size of the precursor mass windows (i.e., 25 m/z) to allow less complex MS2 spectra. A chromatogram can be retained over the retention time gradient by cycling over a given number of mass windows in around one second. These are used

for identification and quantification at MS1 and MS2 level (Fig. 2.8). The corresponding chromatograms matching shape and retention time from one mass window can be grouped into a peak group.

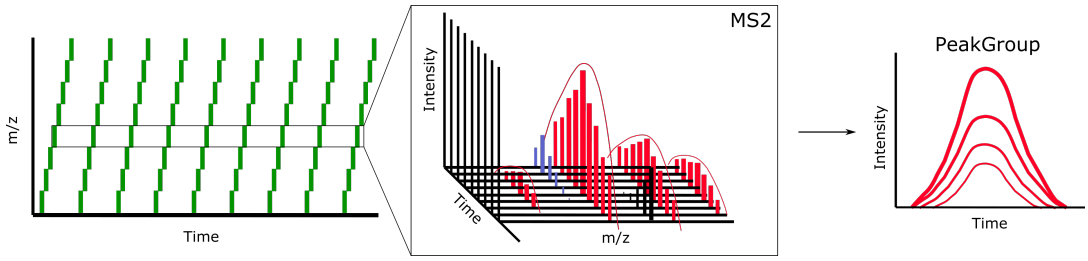


**Figure 2.8: SWATH** All precursors in a specific mass window (e.g., 25 m/z) are fragmented iteratively approximately every second (cycling). This reduces the complexity of the MS2 spectra. Which allows the extraction of chromatograms over the retention time. The overlay of these extracted chromatograms is represented as a peak group. These are used for the identification and quantification of an analyte.

## 2.3 Computational Mass Spectrometry

Computational mass spectrometry is a research field dealing with the development of algorithms and software for the analysis of high-throughput experiments related to MS.

### 2.3.1 Centroiding

Detectors used in mass spectrometers record a profile spectrum, which is a continuous signal of m/z ratios. Here, a signal of a specific ion is distributed around the ions actual m/z value. This peak can be converted into a single m/z value by integrating the profile peak abundances. This computational method is called centroiding or peak picking. In some cases, the vendor software centroids (a part) of the data automatically, but in most cases, this has to be performed after the data acquisition using open-source software. The actual use of centroided or profile data depends on the algorithms of later analysis steps.

### 2.3.2 Quantification (DDA)

MS-based metabolomics aims to quantify all metabolites in a sample. Quantification is possible at MS1 and MS2 levels. The level used depends on the acquisition method and analysis software. In DDA experiments, quantification of analytes uses the MS1 level by integrating the signal of the MS1 monoisotopic peak over the retention time of the analyte. The quantification can be either absolute or relative. Absolute quantification is possible by using standard substances. Since the concentration of the standard is known, the exact quantity can be calculated based

on the measured intensity of a similar analyte. Relative quantification is performed when comparing the same compound detected in a case and control in the same experiment.

### Feature Detection

The first step in quantification is feature detection. It aims to detect possible analytes - features - in the MS map. A feature is defined by the parameters m/z, retention time, charge, and intensity. It represents all signals from one analyte with a specific charge (Fig. 2.9).

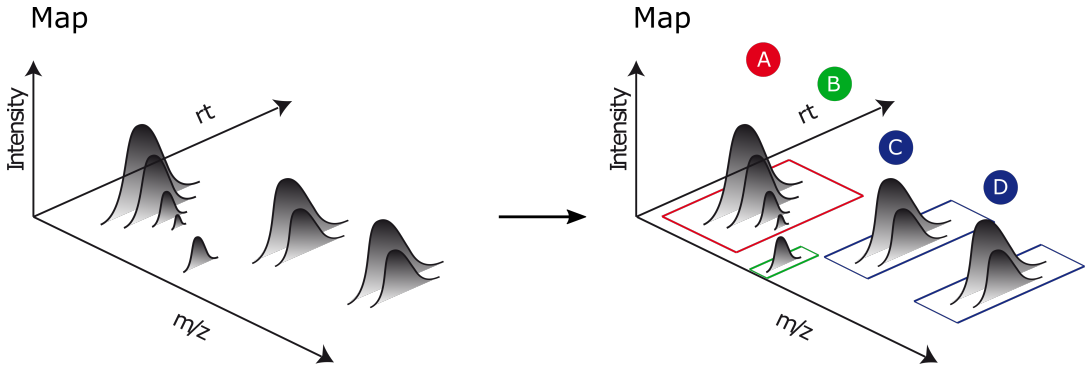


**Figure 2.9: Feature Detection** It aims at the detection of possible analytes defined by an m/z, retention time, charge and intensity. In this example four individual features were identified in the MS map (A, B, C, D).

### Adduct Grouping

In the next step, the detected features are screened for the availability of adduct ions. An adduct ion describes a charged ion, that is formed by interaction of two species, usually an ion and a parent molecule [M] to form an ion, containing all the constituent atoms of one species and an additional atom or atoms[29]. An analyte can form different adduct ions in the ESI, most common is the protonated form of the parent molecule [M+H]+ in positive ion mode[30]. Due to this adduct formation, the same potential analyte can be present with varying forms of adducts, e.g., [M+H]+, [M+Na]+, [M+K]+. Therefore, adduct grouping assesses the same analyte with different adduct species in the sample by using the adduct mass difference (Fig. 2.10).

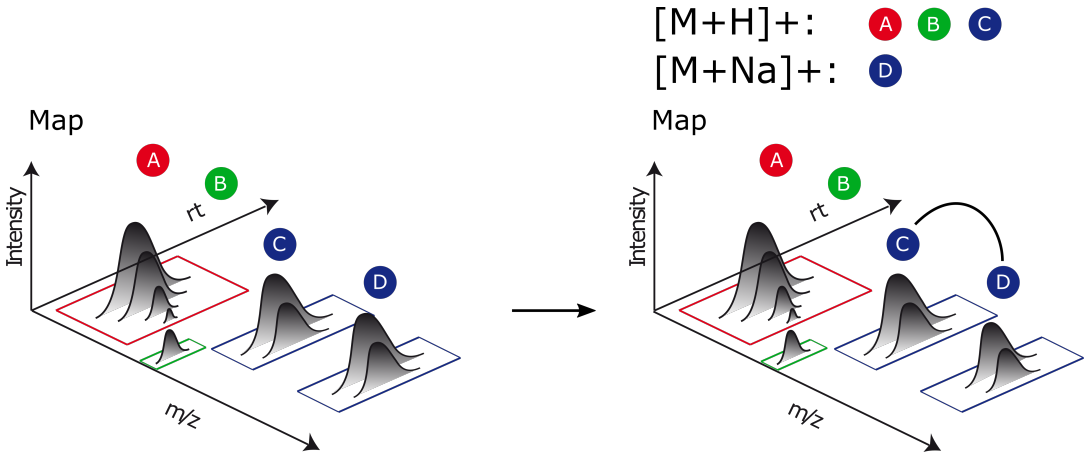


**Figure 2.10: Adduct Grouping** An analyte can be present in different adduct ion forms e.g., [M+H]+, [M+Na]+, [M+K]+. Here, adduct grouping assesses the same analyte with different adduct species in the sample. Features A and B are individual features with an annotated [M+H]+ adduct. In contrast, C and D may stem from the same analyte with different adducts forms [M+H]+ (C), [M+Na]+ (D).

### Feature Linking

Feature linking allows the aggregation of feature information over multiple samples. For example, after processing numerous samples, features showing only a small m/z (e.g., 10 ppm) and retention time error over different MS maps are linked to a single consensus feature (Fig. 2.11). The quantitative values for a specific analyte from different MS maps are assigned to this consensus feature. This aggregation allows easier post-processing. For example the linked quantitative values from the same feature over different conditions can be compared. The actual post-processing is dependent on the aim and the experimental design of the study. In the case of technical or biological replicates, the mean and standard deviation can be calculated.

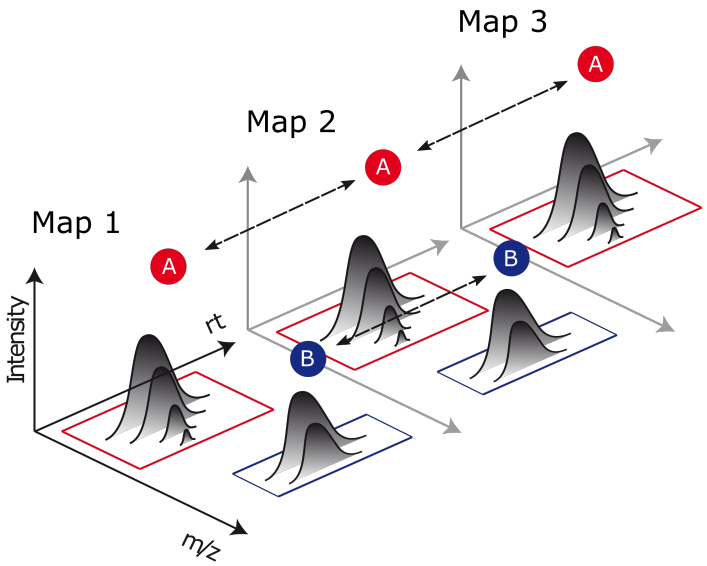


**Figure 2.11: Feature Linking** Features showing a small m/z and retention time deviation over different MS maps are linked. Feature A can be detected and linked in three samples (maps). Features of analyte B were found and linked between two maps.

### 2.3.3 Quantification (DIA)

The quantification of analytes detected in DIA data can be performed via targeted extraction based on MS1 and MS2 signals. Here, the user provides a so-called assay library consisting of prior knowledge regarding the analytes (e.g., mass, retention time). Usually, a precursor and a fragment pair are stored as a so-called transition. This transition is then extracted from the DIA data, quantified, and scored.

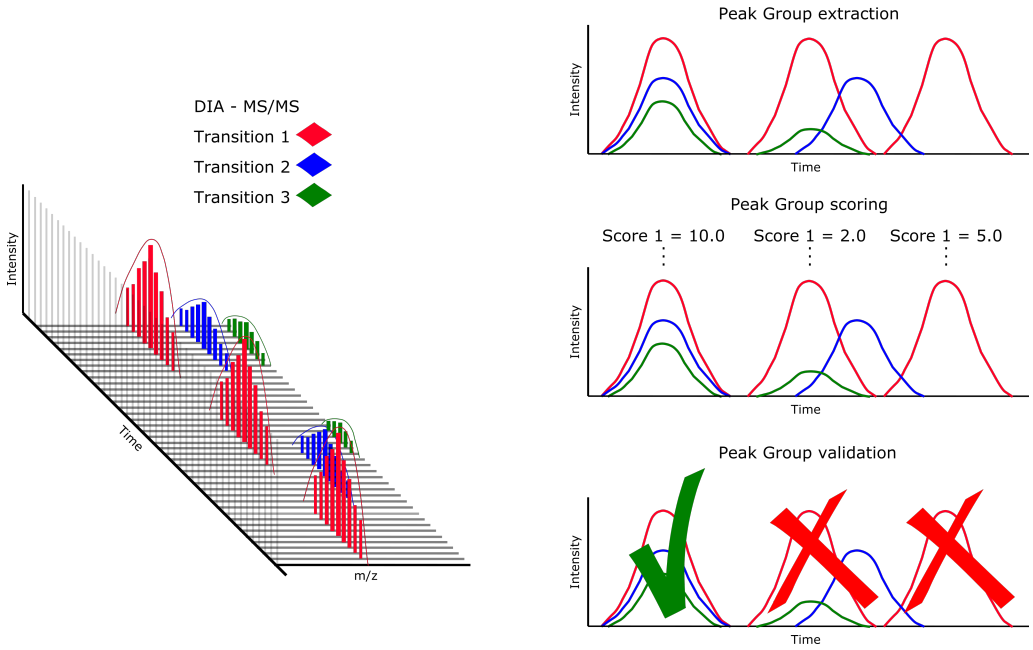


**Figure 2.12: Targeted Extraction and Scoring** Three transitions belonging to an analyte are extracted from DIA MS/MS data in a defined retention time window. Peak group candidates are extracted as chromatograms, scored and validated.

For one assay library entry, defined by m/z and retention time of its transitions, multiple peak groups candidates can exist in the retention time window used for targeted extraction. The filtering of false-positive metabolic peak groups in the low signal-to-noise ranges of DIA results is currently a problem in the field[31]. Either a time-consuming manual validation or a peak group FDR is needed to decide if the quantified peak group is correct (true positive) (Fig. 2.12).

### 2.3.4 Identification

Apart from quantification, another objective of metabolomics is identifying analytes. Several identification methods will be discussed below.

#### Accurate mass search

Accurate mass search is based on the experimental m/z of an MS1 precursor peak. An analyte database (e.g., Human Metabolome Database (HMDB)[32], PubChem[33], LIPIDMAPS[34]) is searched based on the measured neutral/accurate mass of the precursor ion. The neutral precursor mass will be matched with masses of analytes in a set mass error window, for example, 10 ppm. This identification method can lead to ambiguous hits since many possible biochemical analytes or structures have similar masses, depending on their chemical compositions. For example, searching for a molecular mass of 181.0710 u (including an [M+H]+ adduct) using HMDB will lead to two potential compounds within a 10 ppm mass error ($C_6H_{12}O_6$,

181.0707 u; $C_7H_8N_4O_2$, 181.0720 u). Depending on the adduct search space used, the results may be even more ambiguous.

#### Spectral library search

A spectral library search can be performed for identification. Here, two spectra are compared via a calculated cosine-similarity based on the m/z of their peaks[35]. The higher the similarity, the higher the probability of the same spectrum. One drawback is that spectra from different instruments and different collision energies are hardly comparable. At least, there are many resources with MS/MS spectra available (e.g., NIST[36], Global Natural Products Social Molecular Networking (GNPS)[37], HMDB).

#### Fragmentation tree-based methods

Several fragmentation tree-based methods are available to determine the identification of a compound without the need for any prior knowledge. Two distinct vantage points in terms of these methods exist. The first method uses *in silico* fragmentation, focusing on available structures. Here, an accurate mass search is performed to query one or multiple databases (e.g., Kyoto Encyclopedia of Genes and Genomes[38,39], PubChem, Chemspider[40]) to assess possible candidates. The candidates and their structures are then fragmented *in silico* and later used to compare the theoretical spectrum to the experimental spectrum (spectral matching)[41]. The other method uses the MS1 spectrum information to assess possible molecular formulas. It does this by decomposing the isotope pattern of the analyte[42]. The isotopic pattern is introduced due to the abundance of naturally occurring isotopes, so it is possible to generate and match theoretical isotope patterns with experimental ones, to assess the molecular formula of the analyte. The identified formulas are then used as starting points for the fragmentation trees. The trees try to model the possible fragmentation process based on the detected peaks[43,44]. So basically, the MS2 spectrum is used to identify fragments compatible with the molecular formula[45].

### 2.3.5 False-Discovery Rate

FDR estimation gives an indication of the number of false positives in the analysis. In 2007, Elias et al. introduced the concept of target-decoy FDR in the proteomics field[46]. It is used to distinguish correct from incorrect peptide identification assignments based on a list of target and decoy identifications from a target and decoy database for a specific experimental spectrum. Decoys are generated by shuffling or randomizing the peptide sequences in the database. The assumption made here is that the actual decoy is not in the data set but would be similar to the targets in terms of amino acid composition, peptide length, and mass range. This means a hit

against a decoy peptide has to be a false positive identification and it is assumed, that matches to decoy peptide sequences and false matches to sequences from the target database follow the same distribution. The FDR can be calculated based on the target-decoy search results by dividing the number of decoys by number of targets[47]. Unfortunately, creating plausible decoys in metabolomics is not as straightforward since small molecules usually have diverse structural isomers. In metabolomics, target-decoy-based FDR estimation was only introduced recently for large-scale untargeted metabolomics[1,2]. The fragmentation tree-based method introduced by Scheubert et al.[1] ensures the consistency of the decoy spectra by using fragmentation tree re-rooting. First, a fragmentation tree is constructed for the original spectrum identification assigning fragments compatible with the metabolite substructures. In a second step, this fragmentation tree will be re-rooted. A new root is chosen, leading to tree rearrangements by shifting the fragmentation reaction order. This results in new potential fragments based on the original metabolite. The generated decoys should have a slightly lower probability of occurring in the sample but could potentially represent the same metabolite. The FDR for hits in a spectral library database can be estimated with this method. In the targeted setting, the peak group FDR estimation works differently. In proteomics, experimental specific targets and decoys are added to the assay library (prior knowledge database), which is used for targeted extraction. For available targets and decoys, peak groups are extracted and scored. A discriminant score (d-score) distribution is computed using a linear discriminant analysis based on the available subscores, and statistical error estimates are derived by fitting a null distribution[48,49,50] (Fig. 2.13). Currently, this method is not available in metabolomics due to difficulties with the experimental specific decoy generation.

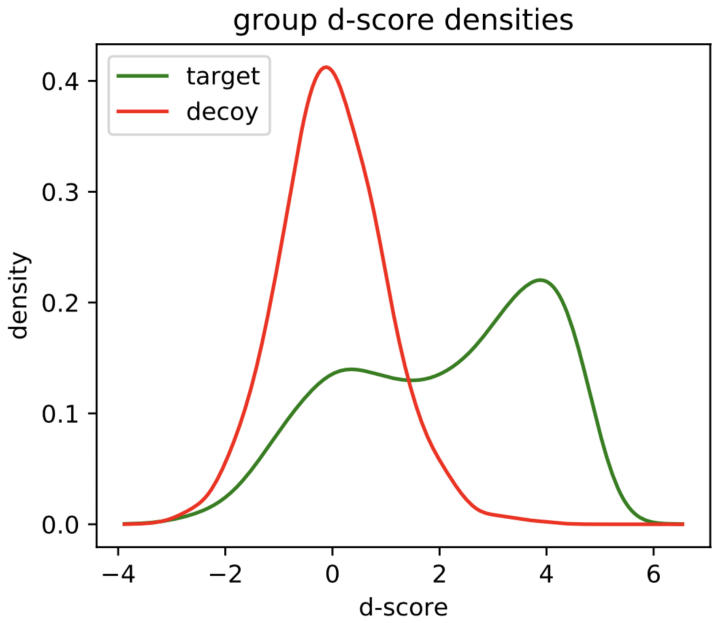


**Figure 2.13: Peak Group d-score Density Diagram** The diagram shows the d-score density distribution for the target and decoy peak groups.

In terms of FDR control, the importance or span of the FDR depends on the research question. For example, a low number of false positives in clinical biomarker identification is important (1% - 5% FDR). On the other hand, if the aim is to find new potential biomarkers, an FDR of 10% might still be valid since these must be further validated.

### 2.3.6 Computational Framework for Metabolomics MS - OpenMS

The efficient analysis of high-throughput MS data in metabolomics is dependent on the availability of a computational framework. This allows the user to perform the necessary analysis steps, such as identification, quantification, and FDR estimation. OpenMS is such a framework[7]. It is an open-source C++ framework, allowing for high-throughput processing of mass spectrometry data (Fig 2.14). OpenMS builds on a few external C++ libraries, using existing high-performance code for specialized actions, such as XML-parsing, graphical user interface (GUI) programming, linear algebra, geometric tools, machine learning, and solvers for integer linear programming. The OpenMS core library contains over 1,300 classes representing concepts of mass spectrometry, consisting of data structures, algorithms, network, and file input/output (IO) for MS data processing. On top of the library, the OpenMS Proteomics Pipeline (TOPP) tools, small applications applicable for a specific task, such as quantification, identification, filtering, FDR estimation, visualization, are built. OpenMS has over 180 applications to analyze bottom-up proteomics, top-down proteomics, metabolomics, RNA-Omics, and cross-linking MS. The library and its algorithms are wrapped for python and R to allow

fast prototyping and method development. In addition, OpenMS is integrated into several workflow systems, such as Konstanz Information Miner (KNIME)[51,52] and Galaxy[53], allowing for user-friendly workflow building and processing.

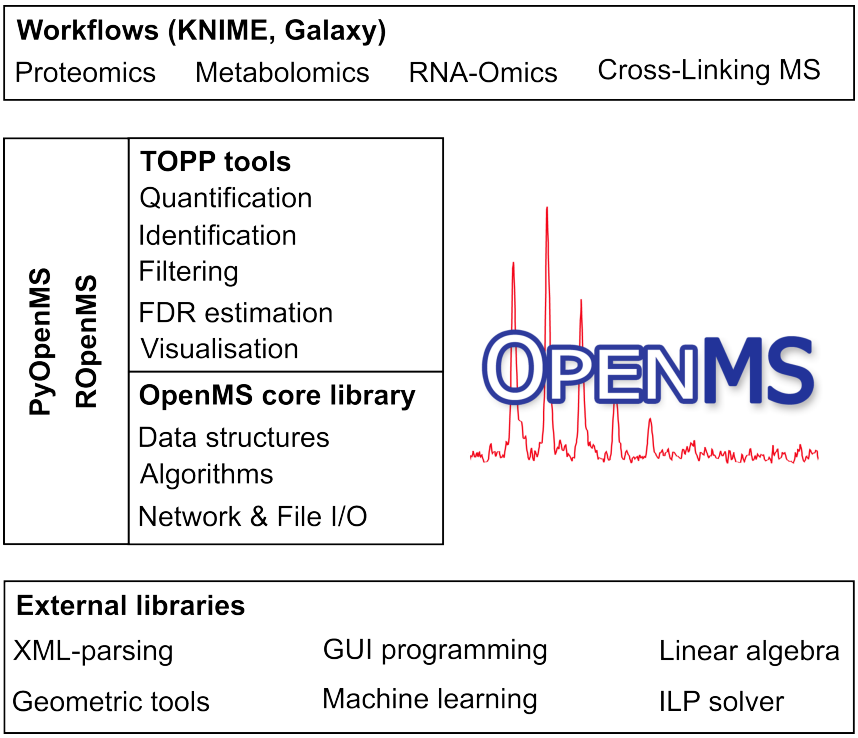


**Figure 2.14: OpenMS framework** The OpenMS core library is dependent on external libraries which provide efficient methods for a specific functionality. The core library itself provides data structures, algorithms, networks, and file input or output for concepts of mass spectrometry. The OpenMS Proteomics Pipeline (TOPP) tools are applications built based on the core library. Each tool has a specific functionality. Multiple tools can be combined into powerful workflows, analyzing all kinds of mass spectrometry -omics data, using the workflow engines KNIME and Galaxy.

### 2.3.7 Role of OpenMS in this Thesis

During his Ph.D. studies, the author has been an OpenMS core developer responsible for maintenance, coordination of development efforts, and release management. In addition, the author developed the tools *SiriusAdapter* and *AssayGeneratorMetabo* used in this thesis, and code to export *MzTab-M* files were integrated into OpenMS and are available from release version 2.8.0 (https://github.com/OpenMS/OpenMS/releases/tag/Release2.8.0).

# Chapter 3

# Automated, FDR Controlled Targeted Analysis for Data-Independent Acquisition in Metabolomics

This chapter includes partially identical or adapted content with permission from:

---

*DIAMetAlyzer: Automated, false-discovery rate controlled analysis for data-independent acquisition in metabolomics*

Oliver Alka, Premy Shanthamoorthy, Michael Witting, Karin Kleigrewe, Oliver Kohlbacher, Hannes L Röst

*Nat Commun 13, 1347 (2022)*

---

A detailed description of the contributions to the project by coauthors is provided in the Appendix C

## 3.1 Introduction

### 3.1.1 Motivation

MS is a flexible tool that allows data acquisition in an untargeted and targeted fashion. While the untargeted approach aims at detecting as many metabolites as possible, the targeted approach focuses on the most accurate quantification of a small subset of metabolites. Thus, targeted methods such as Multiple Reaction Monitoring (MRM) or Parallel Reaction Monitoring (PRM) are limited in analyte coverage but more precise in quantification. Untargeted

approaches often use DDA for fragmentation, in which as many metabolites as possible are recorded in a data-driven (partially stochastic) manner. While this sampling allows for the detection of more compounds, its stochastic nature results in lower reproducibility, an increased fraction of missing data, and reduced quantification accuracy[27,28]. DIA is a novel acquisition method that combines the advantages of targeted and untargeted approaches. It cycles through a series of predetermined mass ranges (DIA or SWATH windows) to acquire a high-resolution MS2 spectrum, thus circumventing the restriction of only sampling a subset of analytes and boosting reproducibility by sampling the entire mass range. This allows for the systematic, unbiased acquisition of fragmentation spectra at the cost of acquiring highly multiplexed spectra since the mass isolation range for each DIA window is generally larger than for other methods. A comparison of DDA and DIA data acquisition revealed that DDA excels in MS2 spectrum quality, whereas DIA shows a better performance in quantitative precision and MS2 spectrum coverage[31]. A major challenge of DIA for the field is the measurement of multiplexed spectra, which are considered lower quality. Two distinct strategies exist to analyze DIA metabolomics data. Most of the current algorithms use an untargeted strategy based on deconvolution and either specialize in identification via spectral library search or in quantification via targeted extraction based on their deconvoluted pseudo-MS2 spectra[54,55,56]. In a targeted analysis strategy, the compounds to be quantified are defined in advance. This requires knowledge of suitable analyte assays, i.e., retention times, and precursor masses with corresponding fragment masses (transitions). These transitions are collected in a so-called assay library which is used to produce fragment-level extracted ion chromatograms (XICs) for each analyte fragment ion around the expected chromatographic retention time. These XICs (one for each fragment ion) have to be verified for quality and compared with an internal (spiked-in) or external standard which is currently a manual and laborious task, requiring specialized expertise and training. While both the creation of assay libraries for DIA analysis and the processing of XICs has been automated in other fields[57], this is not the case in metabolomics. Additionally, a main challenge in targeted metabolomics is the detection of false positive metabolic features in the low signal-to-noise ranges of DIA results that are unable to be filtered[31]. Here, we present a novel workflow based on the targeted strategy which solves all of those issues at once, firstly by integrating a complete end-to-end pipeline including assay library generation into a widely used software suite (OpenMS[7]) and secondly by implementing a novel procedure to estimate robust and accurate false-discovery rates (FDRs) for DIA metabolomics. Our DIAMetAlyzer software combines DDA and DIA metabolomics data by deriving libraries based on high-quality DDA MS2 spectra with few interferences and then subsequently uses DIA to perform quantification, exploiting the improved MS2 coverage and superior quantification performance of DIA[31]. Fully automated construction of the assay library permits the discovery and quantification of new metabolites and still achieves the quantification accuracy of a manually curated targeted

approach. A combination of semi-supervised machine learning and on-the-fly decoy generation permits the estimation of statistically well-calibrated FDRs for the resulting data sets.

### 3.1.2 DIAMetAlyzer Workflow

The workflow utilizes an experiment-specific assay library curated based on available DDA data and is thus tailored to a specific question and instrument. In metabolomics, annotating fragment ions with the underlying structure is not trivially possible, unlike proteomics. We use SIRIUS to annotate fragments using their compositional fragmentation tree approach[44,45]. The method models the fragmentation process based on available MS2 spectra and the chemical composition of the precursor[58]. In proteomics, decoys can then be generated using common approaches to alter the peptide sequence to determine the FDR[46]. Following the idea of a target-decoy approach, we use Passatutto as the basis for re-rooting of fragmentation trees to generate high-quality decoys[1]. The resulting target-decoy assay library allows for the targeted extraction and scoring of targeted transitions from the DIA data with FDR control. The workflow follows multiple steps (Fig. 3.1).

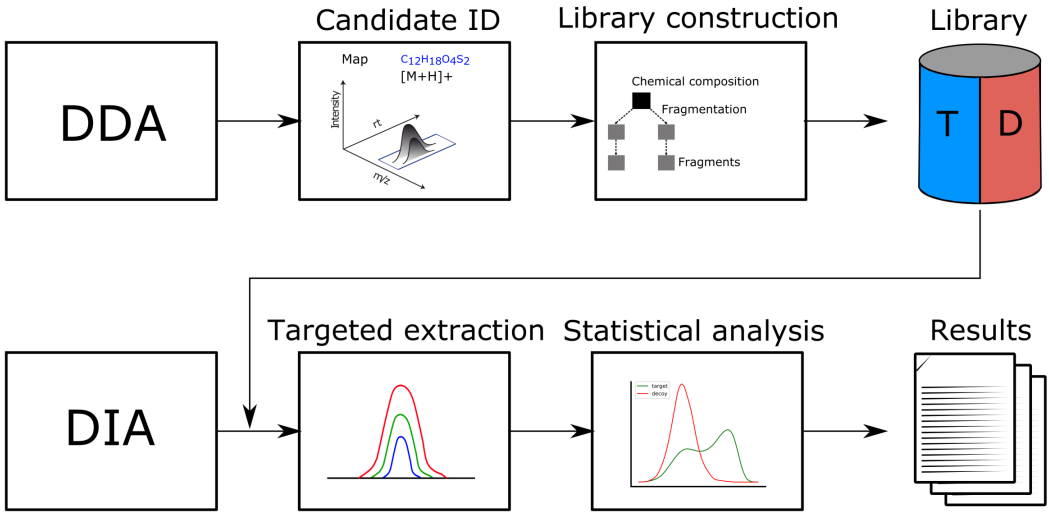


**Figure 3.1: DIAMetAlyzer a Pipeline for Assay Library Generation and Targeted Analysis With Statistical Validation** DDA data is used for candidate identification containing feature detection, adduct grouping, and accurate mass search. Library construction uses fragment annotation via compositional fragmentation trees (SIRIUS) and decoy generation to create a target-decoy assay library using a fragmentation tree re-rooting method (Passatutto). This library is used in a second step to analyze metabolomics DIA data by performing targeted extraction (OpenSWATH), scoring, and statistical validation (PyProphet).

**Candidate identification** For candidate identification feature detection, adduct grouping, and accurate mass search are applied to DDA data.

**Library construction** The knowledge determined about the compound identification, potential adducts, and the corresponding fragment spectra are used to perform fragment annotation via compositional fragmentation trees by using SIRIUS and to extract transitions to build an assay library. FDR estimation is based on the target-decoy approach, with decoys being generated using the recently proposed re-rooting of fragmentation trees by Passatutto, which reduces bias in decoy generation.

**Targeted extraction** The assay library is then used to analyze DIA data. Targeted extraction involves chromatogram extraction and peak group scoring in OpenSWATH[57] (Section 2.3.3, Fig. 2.12). We modified OpenSWATH to support targeted extraction of metabolomics data, included in the latest release of OpenSWATH (see online documentation). Here, chromatograms in a user-specified retention time window are extracted from the DIA data based on the transition entries in the assay library, they encompass precursor and its isotope traces as well as the specified MS2 fragment traces. All extracted traces are grouped in so-called peak groups, which represent a possible analyte with MS1 and MS2 traces. For each peak-group a score matrix is generated based on different scores, such as co-elution and retention time. A detailed description of the OpenSWATH scores can be found in the original publication.

**Statistical Validation** FDR estimation originated from the increasing amounts of data in the genomics field. It is the expected ratio of false positive classifications (false discoveries) to the total number of positive classifications. The "discovery" stands for the items that one labels as "positive", and hence could be true positives or false positives, in the gene expression sense as the genes that you label as differentially expressed[59]. In 2007, Elias et al. introduced the concept of target-decoy FDR in proteomics[46], where it is used to distinguish correct from incorrect peptide identifications. In the targeted field experimental-specific targets and decoys are added to the assay library (prior knowledge database) used for targeted extraction. For available targets and decoys, peak groups are extracted and scored (Section 2.3.3, Fig. 2.12). A discriminant score (d-score) distribution is computed based on the available subscores, and statistical error estimates are derived by fitting a null distribution[48,49,50]. To prevent overfitting, we chose a straightforward linear model (LDA) for target-decoy discrimination using peak group scores with a low cross-correlation, which resulted in an excellent performance on our benchmark data set (Supplementary Fig. A.7).
In terms of FDR control, the importance or span of the FDR depends on the research question. For example, a low number of false positives in clinical biomarker identification is important (1% - 5% FDR). On the other hand, if the aim is to find new potential biomarkers, an FDR of 10% might still be valid since these must be further validated.

## 3.2 Materials and Methods

### 3.2.1 Chemicals

Lyophilized human plasma was obtained from Sigma-Aldrich and prepared according to the supplied instructions (Sigma-Aldrich, Taufkirchen, Germany). LC-MS grade solvents were obtained from Sigma-Aldrich (Sigma-Aldrich, Taufkirchen, Germany). The Agilent LC/MS Pesticide Comprehensive Mix was obtained from Agilent Technologies (Agilent Technologies, Waldbronn, Germany).

### 3.2.2 Sample Preparation

Benchmark samples were prepared by spiking different commercially available pesticide mixes (Agilent Technologies, Waldbronn, Germany) into human plasma metabolite extracts. Human plasma metabolite extracts were prepared by mixing one part human plasma (Sigma-Aldrich, Taufkirchen, Germany) with three parts of precooled acetonitrile (ACN) (4 °C). After centrifugation at 13,000 rpm at 4 °C for 15 minutes, the supernatant was transferred, the solvent was evaporated, and the residue was redissolved in 20% ACN at the original volume of the used plasma aliquot. This matrix was used to dilute the pesticide mixes in a dilution series according to Supplementary Table A.1. Due to the molecular weight range of the pesticide mix, the different steps cover a concentration gradient of 5 orders of magnitude (Supplementary Fig. A.1). For the preparation of the DDA data, each pesticide mix was diluted to 1 ng/µL with either solvent or plasma matrix. For the DIA data, a stock solution of all eight pesticide mixes in the plasma matrix was prepared with a concentration of 1 ng/µL.

### 3.2.3 LC-MS-MS/MS Analysis

The analysis was performed using a Nexera UHPLC system (Shimadzu) coupled to a qTOF mass spectrometer (TripleTOF 6600, AB Sciex). Separation of metabolites from the spiked human plasma metabolite extracts was performed using a UPLC BEH C18 2.1x100, 1.7 µm analytical column (Waters Corp.). The mobile phase was 0.1% formic acid in water (eluent A) and 0.1% formic acid in ACN (eluent B). The gradient profile was 5% B isocratic from 0 to 0.5 min, 5 - 100% B gradient form 0.5 to 10 min, 100% B isocratic from 10 min to 13.5 min, and 5% B isocratic from 13.5 to 16 min. A volume of 5 µL of the sample was injected. As indicated above, different samples were measured in DDA and DIA/SWATH. MS settings were as follows: Gas 1 55, Gas 2 65, Cur 35, Temperature 500 °C, Ion Spray Voltage 5,500 V, declustering potential 80 V Information Dependent Acquisition was used for the generation of assay libraries. The IDA duty cycle was 200 ms for MS1, 80 ms for MS2. The mass range of the TOF MS and MS/MS scans were 50 - 2000 m/z, and the collision energy was ramped from 20 - 50 V or 50 - 80 V, depending on the sample. SWATH acquisition was performed with one TOF MS survey

scan (240 ms) followed by 8 SWATH scans (90 ms). The fragment ion window for SWATH was from 100 to 900 m/z. Here, the used variable window sizes were optimized on the plasma matrix (Supplementary Table A.2).

### 3.2.4 Computational Analysis

The initial DDA data processing for assay library generation was performed as shown previously, using the qTofPeakPicker for centroiding and msconvert for the conversion to mzML[60,61]. The DIA data was converted using msconvert. The data was analyzed using the described workflow with additional manual validation to acquire the ground truth data. Comparisons of ground truth data and additional statistical analysis was performed using Python and R; for further details, see https://github.com/KohlbacherLab/DIAMetAlyzer_additional_code. For a visual inspection of representatives for DDA and DIA data please see Supplementary Fig. A.14.

### 3.2.5 DIAMetAlyzer

Our workflow is composed of steps for candidate identification, library construction, targeted extraction, and statistical validation (Fig. 3.1). **Candidate identification**. Data acquired using DDA is used as input for feature detection, adduct grouping, and accurate mass search. Feature detection is the process of annotating analytes based on their m/z, retention time, intensity, and charge[62]. Based on the feature space, adduct grouping is used to find possible adducts[63]. Annotated features and assigned adducts are then used by accurate mass search to extract potential compositions from a compound database. **Library construction**. Assay library generation is crucial for the targeted analysis of metabolomics DIA data[64]. In this context, we provide a tool called *AssayGeneratorMetabo*. It is implemented using the OpenMS C++ library[65]. The tool uses MS1 and MS2 spectra information and preprocessed feature information to perform precursor correction and filtering based on the number of isotopic traces (data reduction). Afterwards, feature mapping is performed to assign MS2 spectra to a specific feature. To ensure the validity of fragments, fragment annotation assigns fragments to their compatible metabolite substructures. In this approach, SIRIUS4 is used to assign compatible fragments to associated precursors[45]. From the fragment annotated spectra, *n* highest-intensity fragments are automatically extracted to generate potential transitions, which are used for assay library construction. At this stage, a targeted library is available. For the generation of the MS2 decoys, the fragmentation tree-based re-rooting method via Passatutto ensures the consistency of decoy spectra[1], which allows the estimation of a false-discovery rate later in the pipeline. The constructed assay library can be re-used to analyze a multitude of DIA/SWATH samples. Target-decoy assay libraries from other tools could also be used for the next step of the pipeline. **Targeted extraction**. The target-decoy assay library is used to analyze DIA/SWATH data. Targeted extraction involves chromatogram extraction and peak-group scoring. This

step is performed using a metabolomics-extended version of OpenSWATH[57], a well-established workflow commonly used in proteomics, enabling the targeted analysis of DIA data. **Statistical validation**. FDR estimation is performed using the target-decoy library in combination with PyProphet[48,49,50], which was extended to deal with compound information. In PyProphet, the OpenSWATH results were merged, scored on MS1 and MS2 levels using the metabolomics score filter and exported using the export-compound function. The pipeline is available as a KNIME[51,52] workflow using OpenMS[65], pyOpenMS[66], SIRIUS4[45] and Passatutto[1]. The workflow is available in the OpenMS Tutorials (`https://github.com/OpenMS/Tutorials`) and on the OpenMS website (`https://www.openms.de/comp/diametalyzer/`).
An example DIAMetAlyzer KNIME[51,52] workflow is shown in Fig. 3.2. Inputs are the SWATH-MS data in profile mode (.mzML), a path for saving the new target-decoy assay library, the SIRIUS 4.0.1 executable, the DDA data (.mzML), custom libraries and adducts for AccurateMassSearch, the min/max fragment m/z to be able to restrict the mass of the transitions and the path to the PyProphet[49,50] executable. The DDA data is used for feature detection, adduct grouping, accurate mass search and forwarded to the *AssayGeneratorMetabo*. The decoy generation node uses the constructed target library, which will call various python scripts to parse and reformat the input. In addition, Passatutto[1] is called, and the target-decoy assay library is exported. The assay library is used in combination with the SWATH-MS for OpenSWATH[57] and will be automatically analyzed. Pyprophet uses the results for scoring and exports a list of peak groups with associated FDR and quantitative values. For detailed parameters, please check the workflow (20200506_DIAMetAlyzer.knwf - `https://www.ebi.ac.uk/metabolights/MTBLS1108`).

### 3.2.6 Assay Library Generation

For assay library generation, the tool *AssayGeneratorMetabo* was implemented in C++ using the OpenMS Library. It uses spectra information (.mzML) and preprocessed feature information (.featureXML). First, the precursor m/z and the intensity are reannotated, then preprocessing, such as filtering based on the number of isotope traces, can be performed. Afterwards, a feature mapping is used to assign a precursor and its MS2 spectra to a specific feature. After meta-information extraction, fragment annotation is performed using SIRIUS4. From the annotated spectra, $n$ transitions are extracted based on a minimum/maximum intensity threshold. It is possible that the same metabolite and adduct combination is found multiple times in one sample using an accurate mass search. For example, due to experimental reasons, such as column saturation. Currently, the ambiguity is resolved by using the spectrum with the highest precursor intensity. The constructed target library can be exported in various formats (tsv, traML, pqp). Please see the next section for additional details.

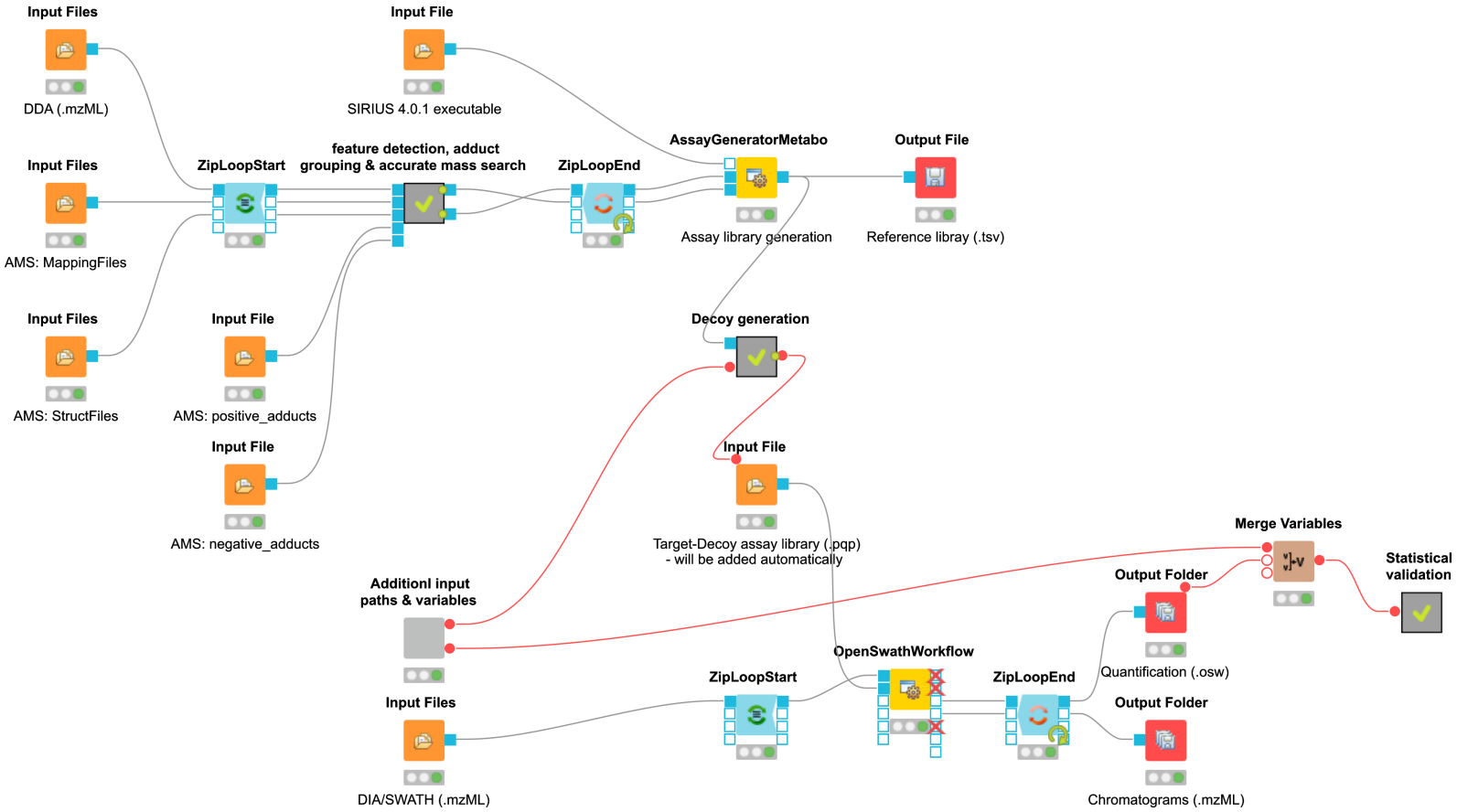


**Figure 3.2: DIAMetAlyzer** Inputs are the SWATH-MS data in profile mode (.mzML), a path for saving the new target-decoy assay library, the SIRIUS 4.0.1 executable, the DDA data (.mzML), custom libraries and adducts for AccurateMassSearch, the min/max fragment mass-to-charge to be able to restrict the mass of the transitions and the path to the PyProphet executable. The DDA is used for feature detection, adduct grouping, accurate mass search and forwarded to the *AssayGeneratorMetabo*. The decoy generation node uses the constructed target library, which calls various python scripts to parse, reformat the input, call Passatutto and export the target-decoy assay library. The target-decoy assay library is processed with the SWATH-MS data in OpenSWATH. PyProphet uses the results for scoring and output a list of metabolites with their respective q-value and quantitative information.

### 3.2.7 AssayGeneratorMetabo Implementation

The *AssayGeneratorMetabo* (*TOPPAssayGeneratorMetabo* class) is implemented as a TOPPTool inheriting from the *OpenMS::TOPPBase* class. The implementation details are presented in two associated flowcharts (Fig. 3.3, 3.4).

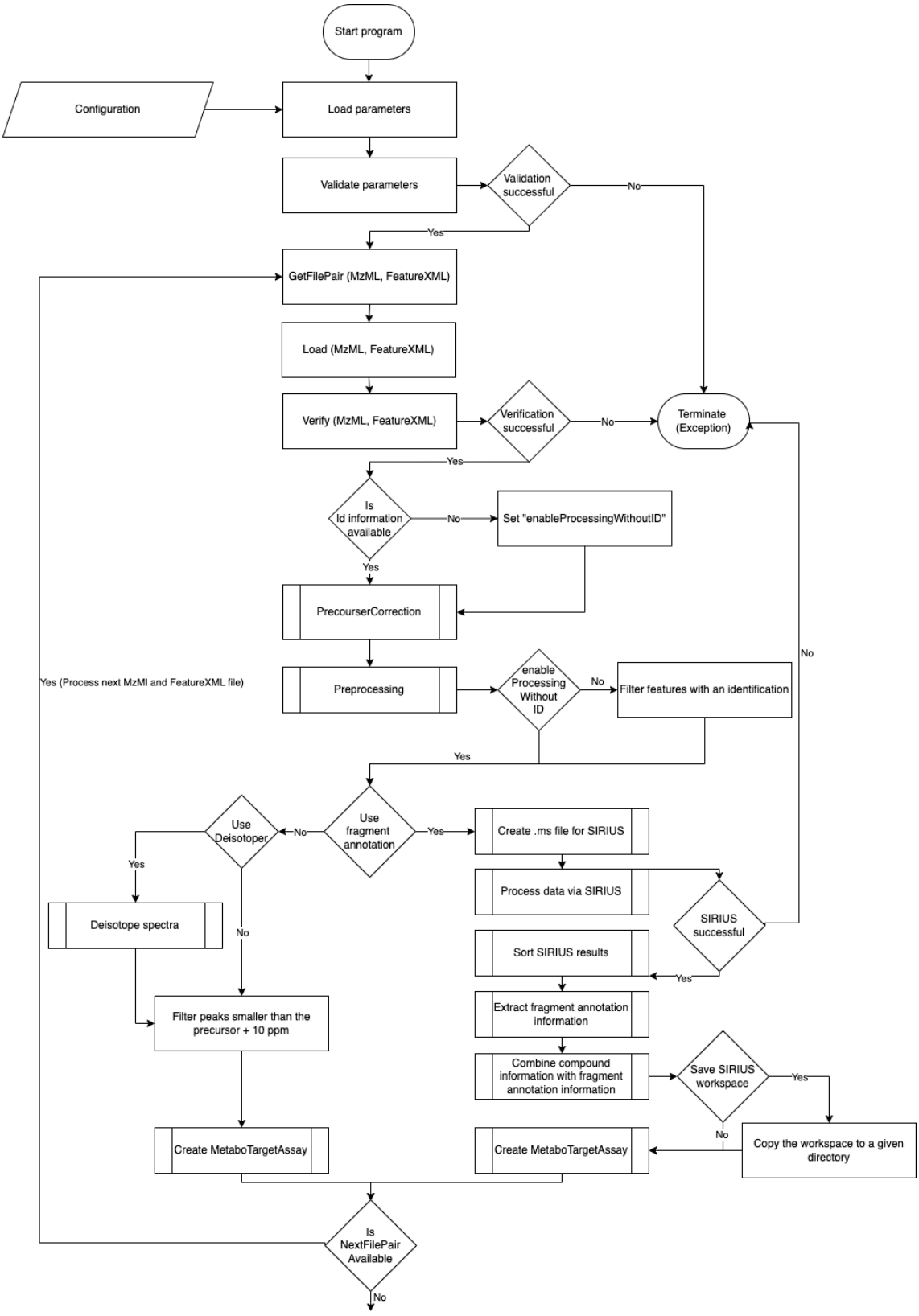
Start program
Configuration
Load parameters
Validate parameters
Validation successful
No
Yes
GetFilePair (MzML, FeatureXML)
Load (MzML, FeatureXML)
Verify (MzML, FeatureXML)
Verification successful
No
Terminate (Exception)
Yes
Is Id information available
No
Set "enableProcessingWithoutID"
Yes
PrecourserCorrection
Preprocessing
enable Processing Without ID
No
Filter features with an identification
No
Yes (Process next MzMI and FeatureXML file)
Yes
Use fragment annotation
No
Yes
Use Deisotoper
Create .ms file for SIRIUS
Yes
Process data via SIRIUS
Deisotope spectra
SIRIUS successful
No
Sort SIRIUS results
Yes
Filter peaks smaller than the precursor + 10 ppm
Extract fragment annotation information
Combine compound information with fragment annotation information
Save SIRIUS workspace
Yes
No
Copy the workspace to a given directory
Create MetaboTargetAssay
Create MetaboTargetAssay
Is NextFilePair Available
No

**Figure 3.3: Flowchart of the AssayGeneratorMetabo Algorithm (Part 1)** After the program startup, the parameters are loaded from the configuration. First, some parameters are validated (e.g., if the SIRIUS executable was provided). If this validation fails, the program terminates. Otherwise, the input MzML and FeatureXML pairs are processed iteratively. After loading the files to memory in the appropriate data structures, additional validations are performed to assess that the same number of mzML and featureXML files were provided and that their primary ms run path matches. Again, the program terminates if the validation fails. Afterward, the featureXML is checked if identification information was provided. If not, the "enableProcessingWithoutID" flag will be set automatically to allow the processing of features without identification. Then precursor correction is performed, and additional preprocessing, such as filtering based on the number of mass traces and mapping of MS2 spectra to features. At this point, only features with identification may be used, depending on the "enableProcessingWithoutID" flag. Two different strategies can be used for the following processing steps, which are dependent on the value of the fragment annotation parameter. If fragment annotation should not be performed and the use of the deisotoper is permitted, the spectra are deisotoped to reduce the number of peaks without reducing the information content. Afterward, peaks smaller than the sum of precursor m/z and 10 $ppm$ are filtered based on the assumption that almost all fragments (substructures) are singly charged and smaller than the precursor. From these, a MetaboTargetedAssay is created, which stores precursor, metadata, and additional compound information. The other strategy uses SIRIUS for fragment annotation. Here, a .ms file is generated as input for SIRIUS. Then the SIRIUS executable with the corresponding parameters is called via a wrapper. If the processing is unsuccessful on SIRIUS side, the program terminates. Otherwise, the results are sorted, and fragment annotation information is extracted and combined with the compound information. If the SIRIUS workspace should be retained, it will be copied to a given directory. Then, a MetaboTargetedAssay is created. Independent of the strategy, the following file pair will be processed (if available).

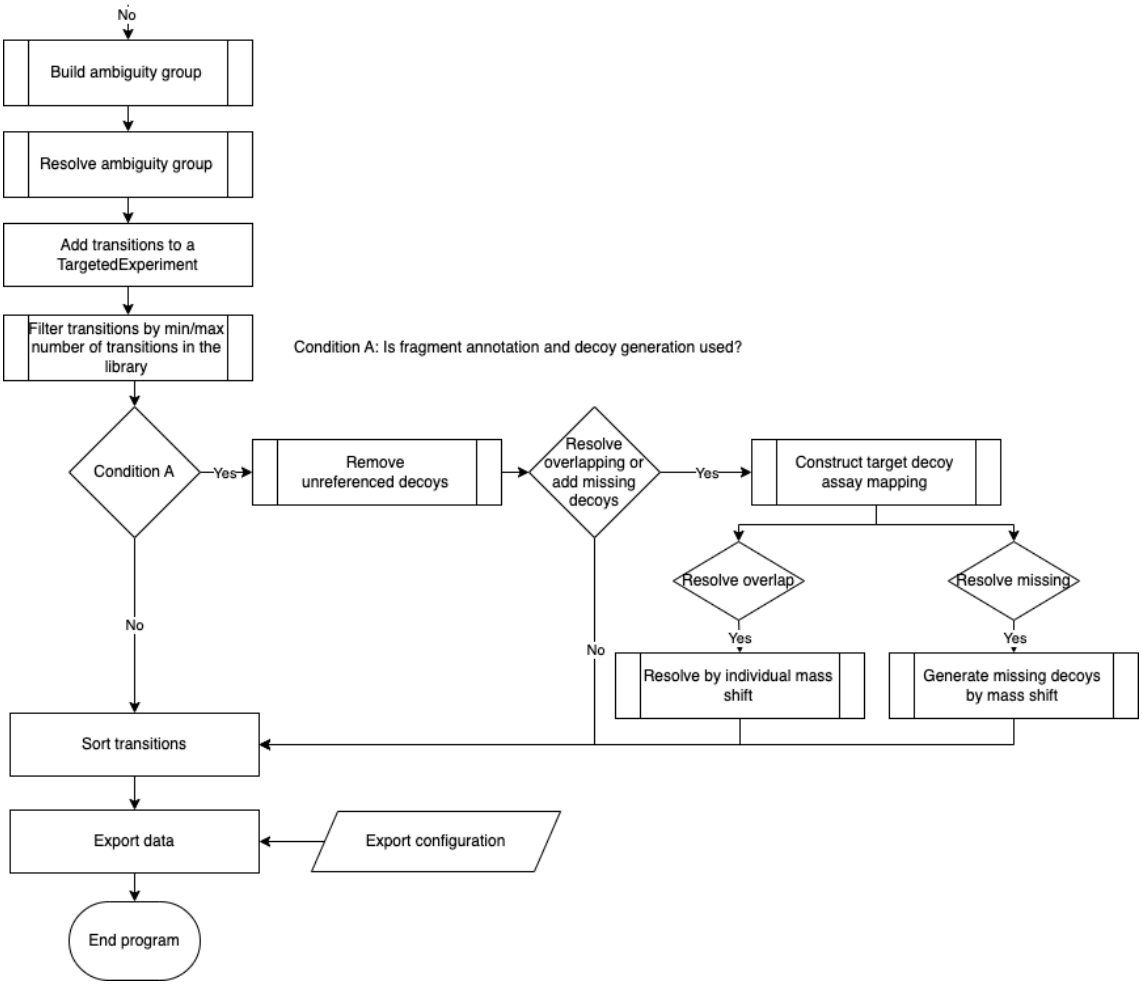


**Figure 3.4: Flowchart of the AssayGeneratorMetabo Algorithm (Part 2)** The list of MetaboTargetedAssay created by processing all mzML and featureXML file pairs is used to build ambiguity groups by performing feature linking over all processed samples. Then, these ambiguity groups are resolved by filtering targets and corresponding decoys (if available) based on sample occurrence (e.g., at least in 20% of the samples). In addition, if multiple possible identifications are reported within one ambiguity group, the one that occurs most often is used. From the resolved groups, the target and decoy with the highest precursor intensity are used for library generation later in the algorithm. Then, the transitions are added to a TargetedExperiment and filtered by the minimum and maximum number of transitions (e.g., minimum three, maximum six transitions). Depending on the usage of fragment annotation and decoy generation, unreferenced decoys are removed, and if wished, overlapping and missing decoys are resolved. Individual mass shifts by the mass of ($-CH_2$) are used to resolve overlapping target and decoy masses. Missing decoys are resolved by creating decoys via a mass shift. Afterward, the transitions are sorted and exported depending on the configuration (.tsv, .traML, .pqp).

After the program starts, the configuration is expected via the command line or a configuration file (.ini). The algorithm loads the parameters and validates some (e.g., should fragment annotation via SIRIUS be performed and was the path to the SIRIUS executable provided via

the configuration). The algorithm will further validate environmental variables and additional commonly used paths if the SIRIUS executable is unavailable. In the case of an unlocatable executable, the processing will terminate with an exception (*Exception::InvalidValue*). Otherwise, the program will evaluate the equality of the number of provided MzML and FeatureXML files. If these do not match, the processing terminates with an exception (*Exception::MissingInformation*). Afterward, the algorithm will continue with a loop over all input files. Here, the MzMLFile and the FeatureXMLFile are loaded into the corresponding data structures (*OpenMS::PeakMap*, *OpenMS::FeatureMap*). The primary ms run path correspondence between the mzML and the featureXML is validated. A failed validation informs the user via a warning. A warning is reported instead of an exception since other tools may overwrite the path while processing the file. The user needs to decide whether to continue the processing. The *AssayGeneratorMetabo* needs centroided spectra for its algorithms. Suppose a profile spectrum is provided, an exception (*Exception::FileEmpty*) is thrown, and the program is terminated. The tool can process featureXML with and without identification information. Features are detected using the *FeatureFinderMetabo*, and the *AccurateMassSearch* can provide identification information. If a featureXML without identification information is used, the *enableProcessingWithoutID* flag is set automatically. At a later step in the algorithm, only features with identification information will be filtered and undergo further processing if the flag is not set. At this point, the verification of the input data and parameters is completed, and further data processing is performed. First, a precursor correction. Here, the precursor m/z and intensity of a given MS2 spectrum is re-annotated via *PrecursorCorrection::correctToHighestIntensityMS1Peak*, by selecting the peak in the corresponding MS1 spectrum with the highest intensity as corrected precursor using retention time and mass range information (e.g., precursor mass $+/-$ 0.01 Da). Afterward, the features are filtered by the number of mass traces, and the MS2 spectra are allocated to a feature within a minimal distance. In the case of multiple features in the tolerance window, the closest one in m/z to the precursor is selected (*SiriusAdapterAlgorithm::preprocessingSirius*). Two distinct strategies are available depending on the usage of fragment annotation via SIRIUS. The first uses an optional deisotoping step (*Deisotoper::deisotopeAndSingleCharge*). This step removes isotope traces from MS2 spectra. Here, peak information is removed without narrowing the overall information content required for the library generation. From a library perspective, it is better to have an additional fragment peak providing new/additional information than a second isotope peak of the same fragment. Afterward, it is filtered for peaks with a smaller m/z than the sum of the precursor m/z and 10 ppm based on the assumption that most fragments are singly charged and that the fragments (substructures) are smaller than the precursor. On the other hand, if fragment annotation is used, a .ms file is generated (*SiriusMSFile::store*). The .ms file is the internal data format used by SIRIUS for processing. Afterward, SIRIUS is called via a QProcess wrapper (*SiriusAdapterAlgorithm::callSiriusQProcess*). If the processing is unsuccessful, an exception is thrown (*Exception::Postcondition*), and the

program is terminated. Otherwise, the corresponding workflow is sorted *SiriusAdapterAlgorithm::sortSiriusWorkspacePathsByScanIndex*. Then the fragment annotations (CSI:FingerID) and decoy annotations (Passatutto) are extracted from the SIRIUS workspace. Ambiguous identifications (multiple identifications for a feature with the same MS2 spectra) are resolved by choosing the identification with higher explained peak intensities (*SiriusFragmentAnnotation::extractAndResolveSiriusAnnotations*). Now the compound information is combined with the annotated spectrum (*MetaboTargetedAssay::pairCompoundWithAnnotatedTDSpectraPairs*). If the SIRIUS workspace should be retained, e.g., for debugging purposes, it will be copied to a given directory. Depending on the chosen processing, a *MetaboTargetedAssay* will be generated, which will hold the potential transitions for one individual file pair (MzML, FeatureXML) (*MetaboTargetedAssay::extractMetaboTargetedAssayFragmentAnnotation*, *MetaboTargetedAssay::extractMetaboTargetedAssay*). Afterward, the information will be stored in a list of *MetaboTargetedAssay*, and the next file pair will be processed. After processing all file pairs, feature linking via the *FeatureGroupingAlgorithmQT* is used to group the features with their annotations/identifications over all samples (MetaboTargetedAssay::buildAmbiguityGroup). As a second step, the identification ambiguity is resolved (*MetaboTargetedAssay::resolveAmbiguityGroup*). Here, targets and corresponding decoys are filtered based on sample occurrence (e.g., at least in 20% of the samples). In addition, if multiple possible identifications are reported within one ambiguity group, use the one that occurs most often. From the resolved groups, the target and decoy with the highest precursor intensity are used for library generation. Afterward resolving the ambiguity groups, the transitions are added to a *TargetedExperiment*, and the assay is filtered by the number of transitions (min/max transitions) (*MRMAssay::filterMinMaxTransitionsCompound*). Depending if fragment annotation and decoy generation are used, decoys are removed which do not have a respective target after min/max transition filtering (*MRMAssay::filterUnreferencedDecoysCompound*). Now depending on the parameter used for the tool, additional actions can be performed on the decoy generation level. Here, a target decoy mapping is constructed (*MetaboTargetedTargetDecoy::constructTargetDecoyMassMapping*), which can be used to resolve overlapping target and decoy masses (*MetaboTargetedTargetDecoy::resolveOverlappingTargetDecoyMassesByIndividualMassShift*) and to generate decoys using a mass shift approach for missing decoys (*MetaboTargetedTargetDecoy::generateMissingDecoysByMassShift*). After sorting the transitions in the *TargetedExperiment*, it can be exported to one of the three output formats (tsv, traML, pqp). Depending on the output format, additional validation of the *TargetedExperiment* is performed before the export. As a side note, since OpenMS 2.7.0, support for SIRIUS 4.9.0 was added, allowing internal decoy generation (via Pasatutto). In addition, the *AssayGeneratorMetabo* was extended to handle targets, decoys, and internal feature linking to resolve ambiguity. Finally, based on the algorithmic and functional improvements, the DIAMetAlyzer KNIME pipeline was simplified

to allow an improvement in its usability. The improved workflow is available at `https://github.com/OpenMS/Tutorials/blob/master/Workflows/DIAMetAlyzer.knwf`.

### 3.2.8 Decoy Generation

The fragmentation tree-based method from Passatutto was used for decoy generation. The fragmentation trees were acquired using fragment annotation via SIRIUS4. The SIRIUS4 tree format had to be parsed into a Passatutto compatible format. After re-rooting, the decoy-spectra were used to extract transitions. A -$CH_2$ mass was added to the overlapping decoy transition for overlapping transition and decoy-transition masses after extraction. If re-rooting of the tree failed or the fragments were similar to the target ones, -$CH_2$ was added to the original fragment masses as a fallback mechanism to ensure the generation of a decoy and to provide the same number of targets and decoys in the assay library. These fallbacks were used in around 13% (similar to targets) and 5% (re-rooting failed) of the cases (Supplementary Fig. A.8,A.9). Afterwards, the *n* highest intensity peaks were extracted to be used in the target-decoy assay library. On MS1, no decoy was generated.

### 3.2.9 Manual Validation

The assay library was converted to a transition list using an in-house script (`https://github.com/KohlbacherLab/MetaboAssayLibToSkylineTransitionListConversion`). The manual validation was performed using Skyline (19.1.0.193) on default settings unless specified differently. The following transition settings were used. Fragments and precursors were used with the adducts ([M+H][M+K][M+Na]), and all matching transitions were automatically selected. The instrument was set to 100 m/z (min) and 900 m/z (max) and retention time from 0 to 16 minutes. MS1 filtering (up to three isotope traces) and MS/MS DIA with custom SWATH windows (Supplementary Table A.2) were used. In addition, only scans within 2 minutes of MS/MS IDs were used.

### 3.2.10 Assessment of the FDR Calibration

We annotated each peak group from our assay library manually. Here, a visual inspection was performed of the peak groups presence, co-elution, and chromatographic shape. A true positive peak group is present if the precursor and transitions are properly co-eluting and show a chromatographic profile and the peak group is aligned within the dilution data set (decreasing intensity along the dilution series). If the peak group was not of high-quality (i.e., noise), it was excluded from the ground truth. Next, the FDR calibration was assessed by comparing the manually validated peak groups with those automatically detected. We constructed a confusion matrix for a predicted FDR threshold from 0.1% to 30% FDR. The confusion matrix reveals

how many true and false hits we have detected based on the ground truth. We report a false positive when our software found a peak group where none was manually annotated or if the retention time deviation was higher than 5s. From the manual annotation, we compute the true false discovery rate: FDR = FP / (FP+TP). Finally, the true FDR was compared to our estimated FDR using DIAMetAlyzer to assess its calibration. In addition, the matrix was used to determine other metrics such as precision and recall.

### 3.2.11 Comparison with MS-DIAL

The comparison between tools was based on the MTBLS1108 data set (`https://www.ebi.ac.uk/metabolights/MTBLS1108`) using MS-DIAL (Version 4.60). The data was preprocessed, and the assay library was converted to a spectral library (.msp) using an in-house script (`https://github.com/KohlbacherLab/MetaboAssayLibToMSPConversion`). Secondly, the PesticideMix SWATH files were converted to .abf using the ABF Converter (`https://www.reifycs.com/AbfConverter/index.html`). The sample with the highest pesticide mix concentration was used for the comparison. The data was processed in MS-DIAL using default parameters, additionally allowing retention time scoring and the adducts [M+H]+, [M+Na]+ and [M+K]+. In addition, the spectral library and the experiment file, including the SWATH windows, were specified (assaylib_20-50_100_ms_dial.MSP; Experiment_file.txt). The MS-DIAL results were preprocessed by filtering for compounds identified via spectral library search based on a reference MS2 and by conversion of the retention time to seconds instead of minutes (convert_and_filter_peak_list.py). Then, the results were compared to the ground truth data set and DIAMetAlyzer results at 5% FDR (comparison_ground_truth_pyprophet_ms-dial.py). The comparison was visualized using an R script (vis_comp_MS-DIAL.Rmd).

### 3.2.12 Comparison with MetaboDIA

The comparison between tools was based on the MTBLS417 data set (`https://www.ebi.ac.uk/metabolights/MTBLS417`) using MetaboDIA (Version 1.3) and DIAMetAlyzer OpenMS development version (14f627e). The data was preprocessed as previously stated[61,67,68,69]. Libraries were generated by both tools, with identification based on accurate mass search using the databases HMDB[32] (4.0) and LIPIDMAPS[34] (092020). All libraries were used for targeted extraction. Furthermore, statistical validation was performed, and the results were reassessed based on chromatographic retention time alignment[70]. Next, features with an FDR of 0.05 and only the top scoring peak group (rank 1) were used for post-processing analysis. The identification of the top significant features was assessed using MASST Search[37] (Supplementary Table A.3). Using MetaboAnalyst[71] (Version 5.0) - over-representation analysis - based on the first compound name annotated by accurate mass search was performed. Here,

the super-, main- and sub-classes were identified for metabolites and lipids. Please see the Supporting Information (A.1.10) for further details.

### 3.2.13 Code Availability

OpenMS as open-source software is distributed under a BSD three-clause license and is available on Github (`https://github.com/OpenMS/OpenMS`). Additional code for re-analysis of the pipeline can be found on Github (`https://github.com/KohlbacherLab/DIAMetAlyzer_additional_code`).

## 3.3 Results

### 3.3.1 FDR Filtering and Library Coverage

To assess FDR estimation accuracy and quantification performance, the developed pipeline was used for assay library generation and the subsequent analysis of the benchmark data set (see Section 3.2). The analysis was performed automatically via DIAMetAlyzer and benchmarked against the manually annotated ground truth extracted via Skyline[72]. Using the DIAMetAlyzer workflow, we were able to reduce the number of false-positive peak groups by 91% (from 1,471 to 125) when applying a 5% FDR threshold to our results (Fig. 3.5a). The number of true positive peak groups was reduced by 12% by the filtering step (from 3,479 to 3,071). Applying a 1% FDR filter, false-positive peak groups were reduced by 98% (from 1,471 to 19), and true-positive peak groups by 28% (from 3,479 to 2,523). This demonstrates that our workflow can reduce the number of false-positive detections/quantifications through an accurate target-decoy based false-discovery rate approach for DIA data analysis in metabolomics.

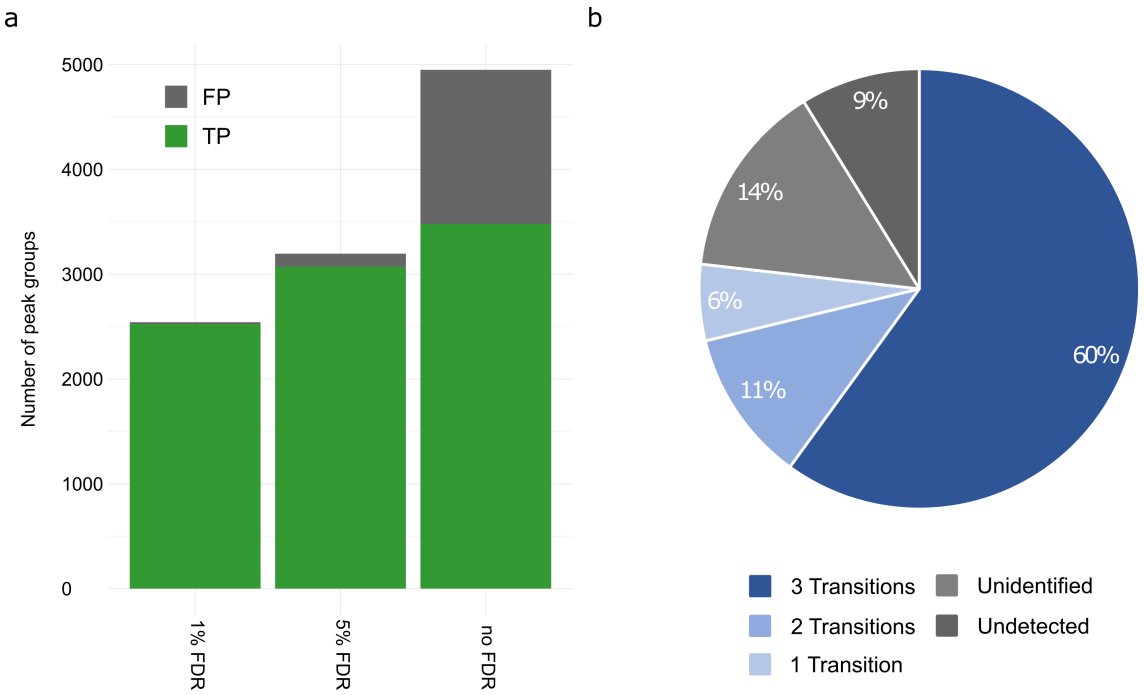


**Figure 3.5: FDR Filtering and Library Coverage** a) Peak groups detected and quantified by the DIAMetAlyzer in an APM spiked-in human blood plasma dilution series (SWATH - 30 samples) filtered by different FDR thresholds. Without FDR filtering (no FDR) we detected and quantified the highest number of true-positive peak groups (n = 3,479), but also the highest number of false-positive peak groups (n = 1,471). At 5% FDR, 3,071 true peak groups and 125 false positives were quantified (3.9%). At 1% FDR the true positive peak groups were further reduced (n = 2,523), so were the false positives (n = 19; 0.7%). b) Individual pesticide mixes in solvent (around 30 pesticides each) were used to construct the target-decoy assay library. Stringent filtering allows high-quality assays to be used in library construction: Around 9% of the pesticides could not be detected in the data. An additional 14% were not identified via MS1 or did not possess a valid MS2 spectrum (4+ peaks, to allow for fragment annotation). 77% of the pesticides were automatically detected, identified, and annotated. In the library construction step, filtering by the number of transitions greatly affects the coverage of metabolites (three transitions: 60% coverage, two transitions: 71% coverage, one transition: 77% coverage).

Following the pipeline from the start, an assay library was generated using reference mixes (Agilent Pesticide Mix, APM) diluted in solvent, then measured using DDA acquisition and finally analyzed using the DIAMetAlyzer workflow (Fig. 3.5b). Since the goal was accurate identification and quantification, only high-quality assays were included in the library. In addition to 9% undetected pesticides, we filtered 14% of compounds that could not be detected via MS1 or did not possess a valid MS2 spectrum (fewer than four peaks to allow for fragment annotation). In the library construction step, filtering by the number of transitions greatly affects the coverage of metabolites (three transitions: 60% coverage, two transitions: 71% coverage, one transition: 77% coverage). By using data from multiple collision energy ranges

(20-50 eV and 50-80 eV), coverage of the assay library can be increased by 11% to 71% (Supplementary Fig. A.3).
To ensure the development of a high-quality assay library, theoretical simulations were used to determine the number of transitions required to reduce ambiguity and improve the number of unique identifications. Using the pesticides data set with the NIST 17 LC/MS library as a combined background metabolome, MS methods were simulated with varying accuracy for both the precursor and fragment m/z windows while alternating the selected number of transitions for each compound. Scoring both MS levels using three transitions increased the number of uniquely identified compounds in our simulation by 2.8-fold and 1.5-fold in comparison to MS1-only and MRM-based analyses respectively, demonstrating the importance of both high-resolution MS1 and MS2 data (Supplementary Fig. A.2). Based on these results, an assay library with three transitions has been chosen for downstream analyses.
The developed assay library was used to perform the analysis of 30 DIA samples with APM spiked-in human blood plasma acquired in DIA mode using SWATH (Supplementary Table A.2). The pesticide mix was measured in triplicates and spiked into human plasma in a 4-fold dilution series, spanning over five orders of magnitude in dynamic range (Supplementary Table A.1, Supplementary Fig. A.1). The data were measured in a ten-step concentration series at two collision energy ranges. The targeted extraction was performed automatically via DIAMetAlyzer and benchmarked against the manually annotated ground truth extracted via Skyline.

### 3.3.2 Accuracy of FDR Estimation

To evaluate the accuracy of our FDR estimates, the automatic and manual analyses were compared to determine the deviation of the ground truth FDR from the FDR estimated by DIAMetAlyzer (Fig. 3.6). We found that the FDR estimated by fragmentation tree re-rooting is slightly conservative, with a slight overestimation for the data acquired at lower ranges of collision energy (20-50 eV) (Fig. 3.6a). In comparison, FDR estimates, for data acquired at higher ranges of collision energy (50-80 eV), demonstrated an increased abundance of overlapping fragments, resulting in more ambiguous analyses (Supplementary Fig. A.4). To assess the accuracy of the classifier, we determined the precision and recall based on different estimated FDR thresholds using the best peak group rank (Fig. 3.6b). Our approach produced an area under the precision-recall curve (AUC) of 0.96, resulting in over 75% recall at 95% precision (or 5% FDR).

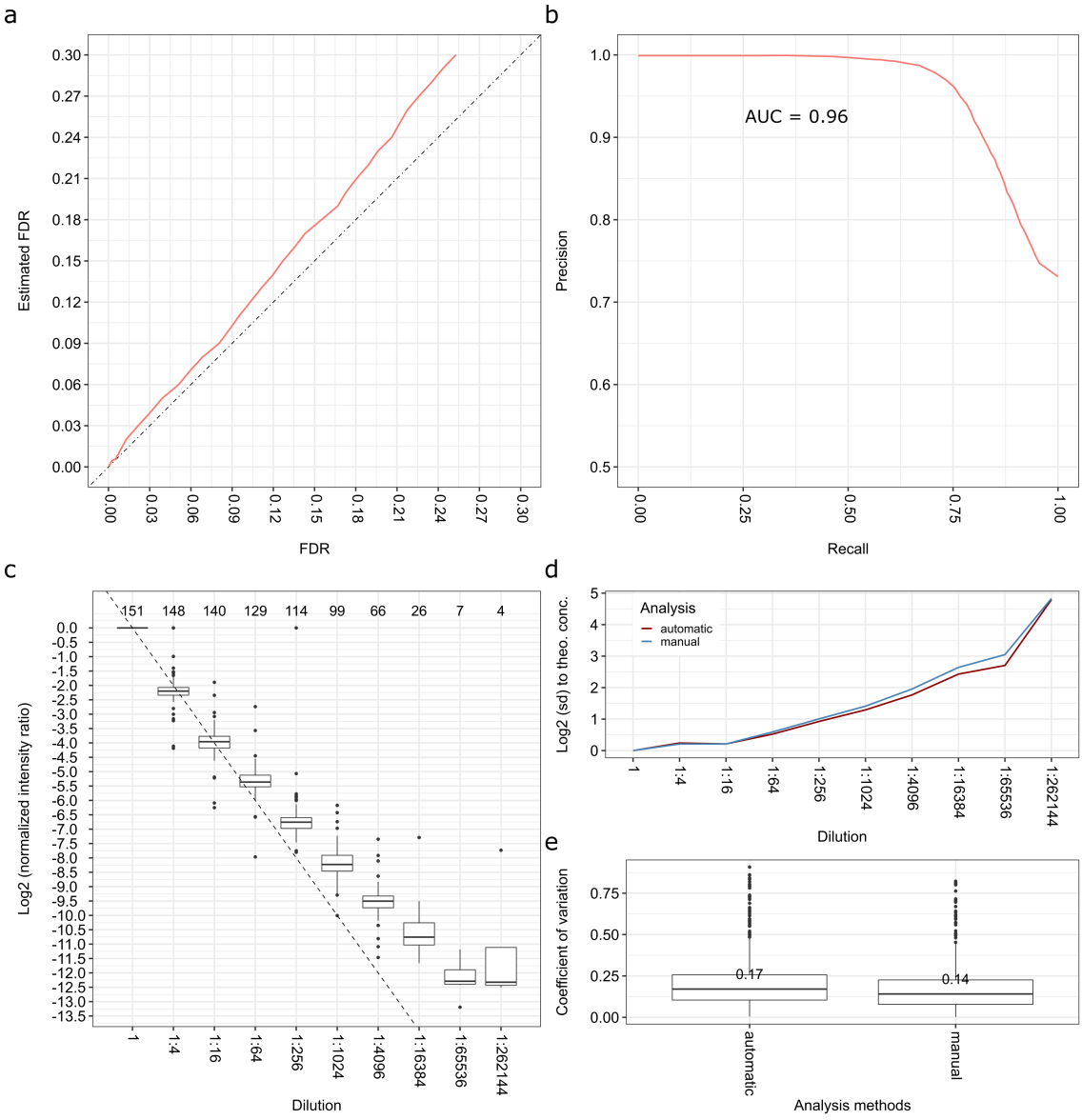


**Figure 3.6: Identification Accuracy and Quantification of DIAMetAlyzer on the Pesticide Spike-in Dataset** a) Estimated FDR versus FDR from ground truth data. b) Precision-Recall curve with the area-under-the-curve (AUC = 0.96). c) Normalised intensity ratio over the dilution series. Dashed line indicates the expected fourfold difference to the next dilution. X-axis (top): The number of metabolites found in the specific dilution at a 5% FDR cutoff. More than half of the initial metabolites could be detected at half of our dilution series (1:1,024). d) Difference in mean, standard deviation in regard to the theoretical concentration of the automatic and manual analysis. e) Median coefficient of variation (CV across three technical replicates for the automatic and manual analysis (CV < 20%)). For c, d, and e, only metabolites detected in triplicates and below a 5% FDR threshold were analyzed, and only true positives were considered in case of panel e.

### 3.3.3 Quantification Performance

To determine the quantification performance, the results were filtered using a 5% FDR threshold and normalized for each combination of metabolite and adduct by the intensity of their highest concentration. More than half of the initial metabolites could be detected at half maximal dilution (1:1,024), based on the last dilution step a metabolite was observed in (Fig. 3.6, Supplementary Fig. A.6). The limit of detection of the individual metabolites was assessed using the unfiltered results, based on an S/N threshold of 10 (Supplementary Table A.4). Comparing the quantification of manual and automatic analyses, the precision of the automated method matches manual analysis and outperforms it in some dilution steps (Fig. 3.6d). In all technical replicates, the median coefficient of variation (CV) of non-normalised quantified signals was smaller than 0.2 (Fig. 3.6e).

### 3.3.4 Comparison to State-of-the-art Algorithms

To benchmark the performance of DIAMetAlyzer against state-of-the-art analysis algorithms, we compared it to MS-DIAL[54] and MetaboDIA[73]. MS-DIAL is a tool specialized in untargeted SWATH analysis based on spectral deconvolution, using computationally constructed pseudo-MS2 spectra for identification via spectral library search. Similar to DIAMetAlyzer, the functionality of MS-DIAL is dependent on the spectral library space provided to the software. Here, to allow a fair comparison between the tools, we used our ground truth APM data set specifying our assay library as a spectral library (Fig. 3.7).

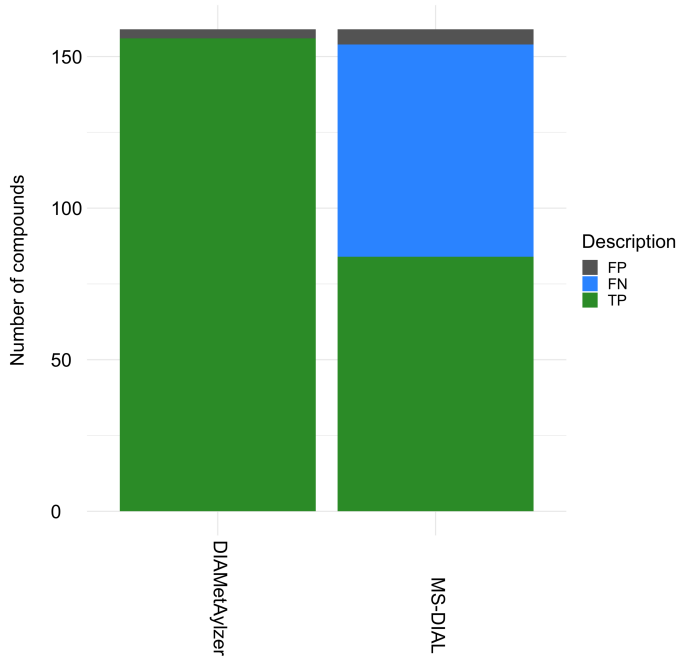


**Figure 3.7: Identification Performance of DIAMetAlyzer in Comparison with MS-DIAL Based on the Generated Assay Library** DIAMetAlyzer was able to identify 156 true-positive and 3 false-positive compounds in comparison to the ground truth. MS-DIAL, with an identification threshold of 0.8 (default) and retention time scoring enabled, was able to identify 84 true positives, 5 false positives and was not able to identify 70 compounds (false negatives).

The DIAMetAlyzer was able to identify 156 true positives and three false-positive compounds in comparison to the ground truth (at 5% FDR). MS-DIAL was able to identify 84 true positives, five false positives and was not able to identify 70 compounds (false negatives). In this setting, we could show the advantage of the DIAMetAylzer targeted extraction strategy with false-discovery rate control based on reference compounds in comparison to untargeted deconvolution. We would like to state that the functionality of MS-DIAL is focused on the identification of unknown compounds and is dependent on the spectral library space given to the software.

MetaboDIA[73], a tool capable of building a consensus MS/MS library based on DDA data using MS-based identification and subsequent quantification via DIA-MS/MS in a non-targeted manner. We used a publicly available age-related macular degeneration (AMD) data set (MetaboLights accession MTBLS417) along with HMDB[32] and LIPIDMAPS[34] for identification via accurate mass to construct a library with each tool. The library was filtered for features found in at least 20% of samples with a minimum of three MS/MS peaks available. Libraries generated by both tools show a significant overlap (66%) of features based on the molecular formula, adducts, and retention time (Fig. 3.8a). DIAMetAlyzer generates a larger number of features

compared to MetaboDIA (46% improvement; 695 compared to 476). Differences between the two libraries are attributable to improved feature detection and more stringent filtering in the assay library creation step in our pipeline.

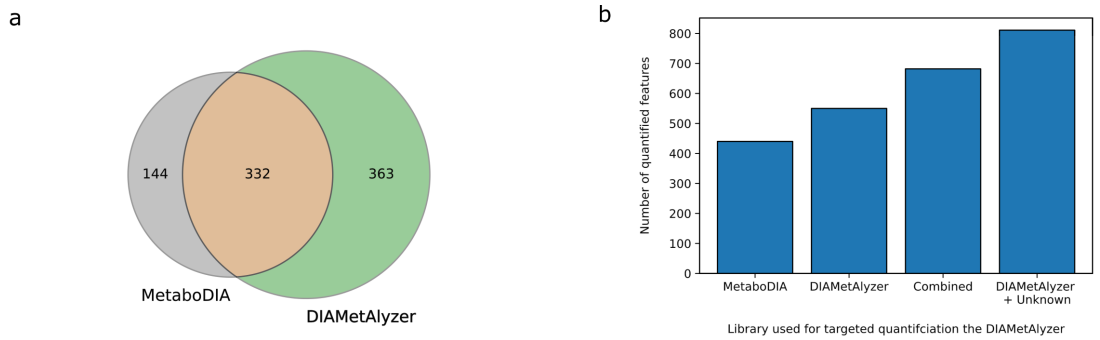


**Figure 3.8: Analysis of Serum Samples of Patients with AMD using MetaboDIA and DIAMetAlyzer** a) Comparison of the library generation of both tools based on features (molecular formula, adduct, and retention time). 66% of the features overlap between the tools. b) Number of quantified features using the various libraries in combination with the targeted extraction of the DIAMetAlyzer. MetaboDIA, DIAMetAlyzer, the library of both tools (Combined), DIAMetAlyzer with the functionality to use known unknowns without prior MS1 identification in addition to the ones with identification (DIAMetAlyzer + Unknown).

Based on the set of molecular formulas and adducts, MetaboDIA was able to quantify 54 features in comparison to DIAMetAlyzer, which was able to quantify 440 features using the same library (MetaboDIA) (Fig. 3.9a). This discrepancy was discussed with the developers of MetaboDIA. MetaboDIA is not in active development. Due to its dependence on XMCS, CAMERA and DIAUmpire, changes in one of these software suites may lead to such performance issues. In addition, based on the overlapping features quantified by both tools, the targeted extraction method used in DIAMetAlyzer has a clear advantage over the method used by MetaboDIA. In most cases DIAMetAlyzer was able to quantify a compound in every sample (Fig. 3.9b).

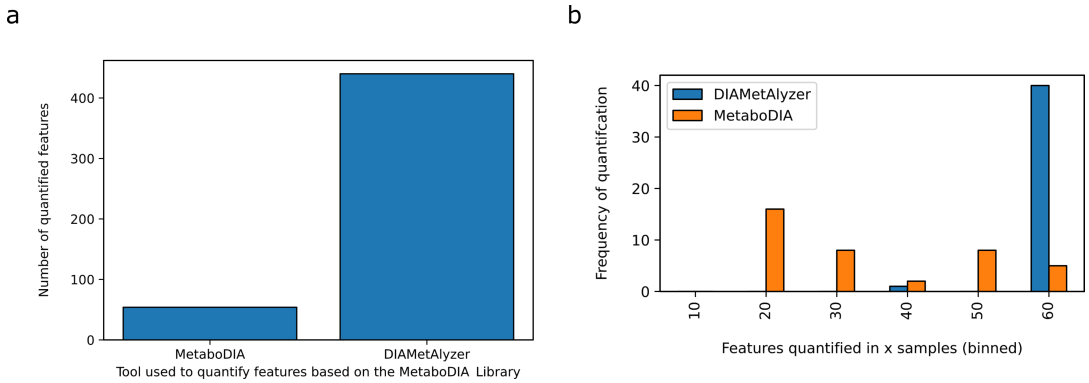


**Figure 3.9: Quantification Comparison Between MetaboDIA and DIAMetAlyzer** a) Based on the set of molecular formula and adduct, MetaboDIA was able to quantify 54 features, in comparison DIAMetAlyzer was able to quantify 440 features using the same library generated via MetaboDIA (MetaboDIA). b) DIAMetAlyzer was able to obtain quantitative values for a feature in around 60 samples. In contrast, MetaboDIA had a higher variation in its quantification frequency.

Using our targeted quantification with the various libraries, we were able to quantify almost twice the features (811 vs. 440) with our library in comparison to the one generated by MetaboDIA (Fig. 3.8b, Supplementary Fig. A.13). When restricting quantification to identified features, DIAMetAlyzer could still quantify 25% more features with its own library than with the one generated via MetaboDIA. Interestingly, there were 144 features uniquely identified by MetaboDIA from the DDA data, allowing us to build a combined library which results in a total of 682 quantified features. These exclusive features were either not detected by our pipeline or were filtered out in the assay library generation step. Additional details on the feature detection, feature linking, and quantitative comparison with MetaboDIA are given in the supplementary material (Supplementary Fig. A.10,A.11,A.12).

To get a general idea of the biological significance of the data, we used the quantified features from the DIAMetAlyzer workflow at a 5% $FDR_{DIAMetAlyzer}$. LIMMA[74] was used with a Benjamini & Hochberg[59] correction for multiple testing to identify differentially expressed features between the conditions control, choroidal neovascularization (CNV), and polypoidal choroidal neovascularization (PCV). We found a total of 118 differentially expressed features with our baseline workflow ($FDR_{Limma} < 0.05$), comparable to the 113 features found using our workflow together with the MetaboDIA library. We were able to report additional differentially expressed features using the combined library (162 features) and found the largest number of differences (220 differentially expressed features) using our identification-free pipeline, almost doubling the number of differentially expressed features.

### 3.3.5 Biomarker Detection

In the following, our aim was to identify individual compounds that could serve as biomarkers of AMD or be involved in disease aetiology. First, we assessed the class separation of the patient groups based on a PCA. Then, we performed a differential expression analysis comparing the individual groups. Last, we performed an evaluation of promising biomarker candidates.

#### Patient Group Comparison

After using our DIAMetAlyzer library (DIAMetAlyzer + Unknown) for the data analysis, a principal component analysis was performed to assess the class separation based on the determined features. Using all available features (Fig. 3.10a) or features which showed a significant difference between groups (5% $FDR_{DIAMetAlyzer}$) (Fig. 3.10b). In both cases, the control shows a distinct group but can not be separated by PC2. The separation in the PC1 direction leads to the conclusion of a batch effect since a few samples from each group (CNV, PCV) are separated by PC1. This occurrence does not arise from the extraction batch or injection order. Unfortunately, not enough metadata is present to explain this effect.

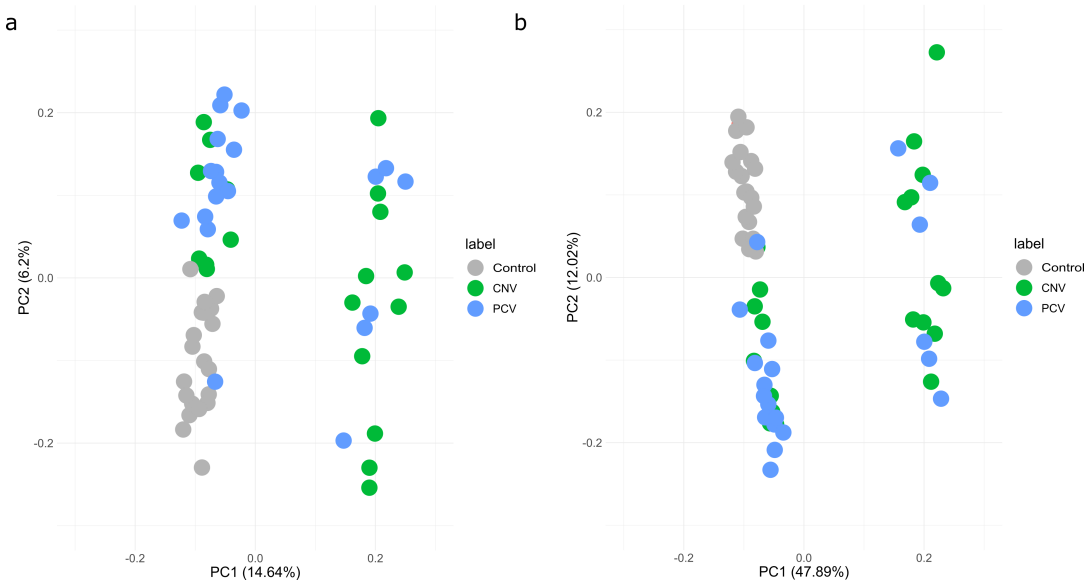


**Figure 3.10: PCA to Assess Group Separation Based on Features** a) PCA based on all available features. b) Based on features which showed a significant difference between groups (5% FDR). In both cases, the control can almost be separated by PC2.

#### Differential Expression Analysis

Differential expression analysis was performed using LIMMA[74] with a Benjamini & Hochberg[59] correction for multiple testing to assess the differences between the individual groups by performing group wise comparisons ($FDR_{Limma} < 0.05$). In the case of control vs CNV, 208

analytes show a significant difference between the groups. 79 analytes in case of control vs PCV. 49 of these analytes are found to be differentially regulated in both groups. In the comparison of CNV and PCV, no significant differentially regulated analytes could be detected. The compound classes of the differentially regulated analytes were retrieved. We found major differences between control and patients in compounds associated with the classes glycerophospholipids, organic heterocyclic compounds, sterol lipids, fatty acids, amino acids, and dipeptides.

**Biomarker Evaluation**

Carnitines and their metabolites are mainly involved in fatty acid metabolism. Oleoylcarnitine ($P_{CNV}$=0.002, $P_{PCV}$=0.01), as well as L-Palmitoylcarnitine ($P_{PCV}$=0.02), are upregulated by around 1.5 times in contrast to the control in both or PCV (Fig. 3.11a,b). These findings follow the research of an altered carnitine shuttle pathway in macular degeneration[75]. Our findings suggest that Linoelaidylcarnitine ($P_{CNV}$=0.04; $P_{PCV}$=0.03), which showed a similar increase in addition to the others, might be a potential biomarker for AMD (Fig. 3.11c). Further possible biomarkers associated with AMD were determined from serum in previous studies, such as phenylalanine, hypoxanthine, tyrosine[76]. We found hypoxanthine levels were significantly increased in CNV ($P_{CNV}$=0.006) by 3.9 times in comparison to the control, which affects the purine nucleotide cycle and can lead to apoptosis of photoreceptors[77,78] (Fig. 3.11d). In addition, gamma-Glutamylphenylalanine ($P_{CNV}$=0.002; $P_{PCV}$=0.0006), gamma-Glutamylisoleucine ($P_{CNV}$=0.01; $P_{PCV}$=0.04) and dityrosine ($P_{CNV}$=0.002; $P_{PCV}$=0.03) were deregulated in both patient groups (Fig. 3.11e,f, 3.12a). Increased serum gamma-glutamyl transferase (GGT) levels were reported as risk factors for AMD[79]. This suggests that gamma-Glutamylphenylalanine, which was detected with an increased intensity by around 1.6 times in comparison to the control, as well as gamma-Glutamylisoleucine (1.7 times increase), could be used as metabolic markers in this case. We found additional biomarker candidates without identification at $m/z$ 601.271468 and retention time 579 s (Unknown A) ($P_{CNV}$=0.0002; PPCV=0.000008) with an intensity increase of 5.2 and 5.7 in comparison to the control and at $m/z$ 944.360964 and 311 s (Unknown B) ($P_{CNV}$=0.005; $P_{PCV}$=0.03) with an intensity increase of 1.7 and 1.9 respectively (Fig. 3.12b,c).

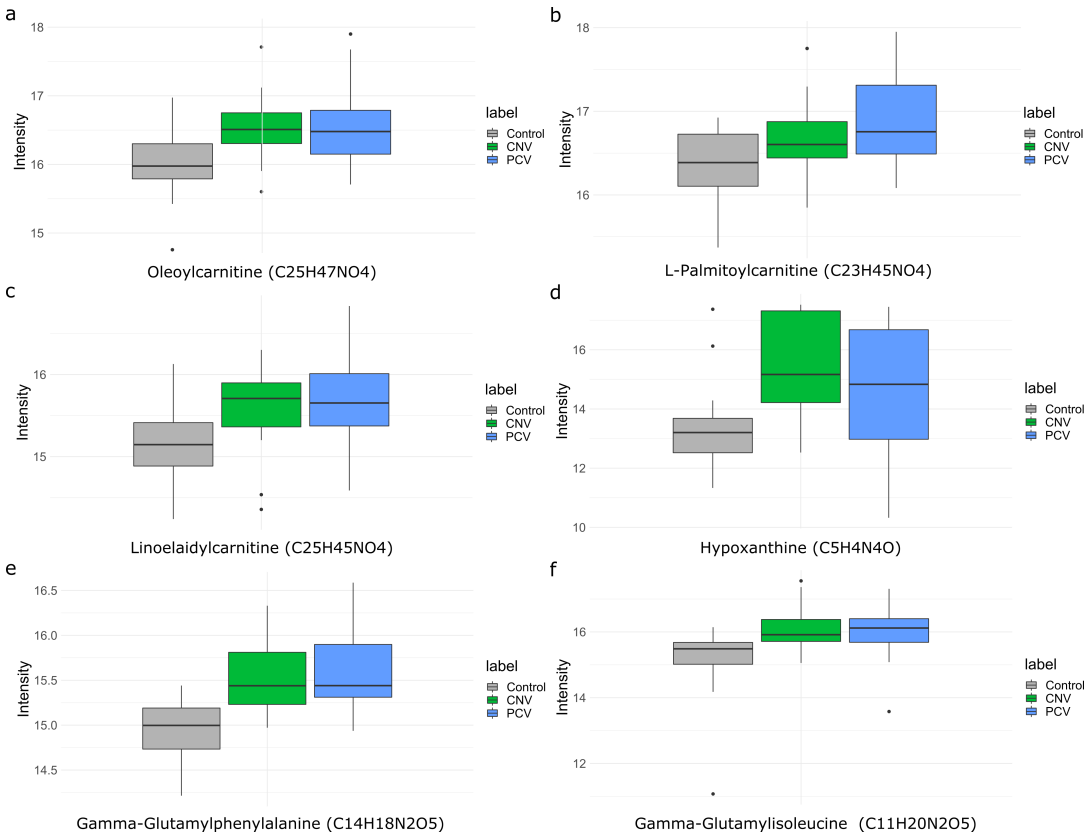


**Figure 3.11: Quantification of Biomarkers and Additional Candidates** a) Oleoylcarnitine ($C_{25}H_{47}NO_4$; $P_{CNV}$=0.002; $P_{PCV}$=0.01) were upregulated by around 1.5 times in both patient groups in contrast to the control group. b) L-Palmitoylcarnitine ($C_{23}H_{45}NO_4$; $P_{PCV}$=0.02) showed the same increase in the case of PCV. c) Linoelaidylcarnitine ($C_{25}H_{45}NO_4$; $P_{CNV}$=0.04; $P_{PCV}$=0.03) showed a similar increase and might be a potential biomarker for AMD. d) Hypoxanthine levels were significantly increased in CNV ($C_5H_4N_4O$; $P_{CNV}$=0.006) by 3.9 times in contrast to the control. e) gamma-Glutamylphenylalanine ($C_{14}H_{18}N_2O_5$; $P_{CNV}$=0.002; $P_{PCV}$=0.0006) with an increase of 1.6 times in contrast to the control. f) gamma-Glutamylisoleucine ($C_{11}H_{20}N_2O_5$; $P_{CNV}$=0.01; $P_{PCV}$=0.04) showed a 1.7 increase in comparison to the mean intensity of the control. The identification of the compounds is based on MS1 accurate mass search and MS2 fragment annotation.

As a validation, dityrosine, which was increased by around 1.7 to 2.4 times in comparison to the control, plays a role in oxidative stress and is associated with macular degeneration[80]. An additional interesting aspect is the significantly deregulated compounds 5,8,11,14-Eicosatetraenoic acid (EPA) ($P_{PCV}$=0.01; $P_{CNV}$=0.04) and 4,7,10,13,16,19-Docosahexaenoic acid (DHA) ($P_{PCV}$=0.006; $P_{CNV}$=0.008) with an increase of 1.4 to 2.0 times in comparison to the control (Fig. 3.12d,e). These have previously been associated with a reduced risk for neovascular AMD[81]. An explanation for this finding in the patients could be an Omega-3 fatty acids rich diet, which is often advised to AMD patients due to their anti-inflammatory properties[82,83].

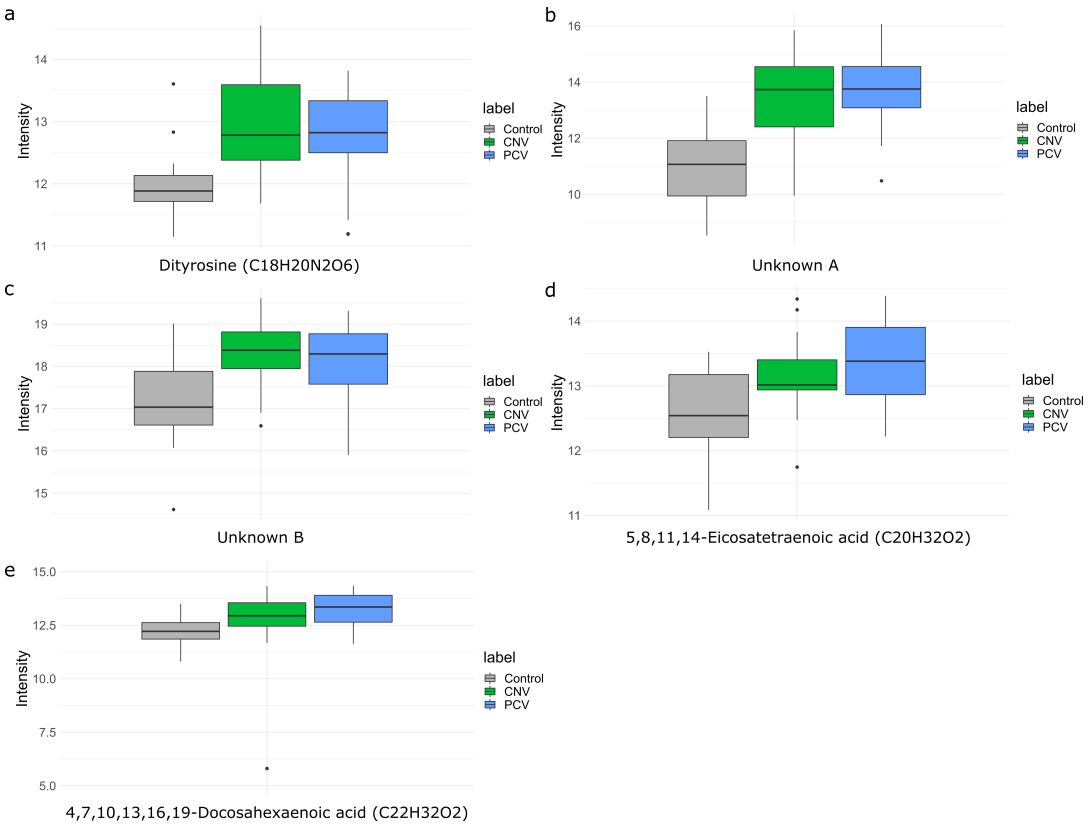


**Figure 3.12: Quantification of Biomarkers and Additional Candidates II** a) Dityrosine, was found to be increased by around 1.7 to 2.4 times in contrast to the control ($C_{18}H_{20}N_2O_6$; $P_{CNV}$=0.002; $P_{PCV}$=0.03). b) Unknown A at 601.271468 *m/z* and a retention time 579 s ($P_{CNV}$=0.0002; $P_{PCV}$=0.000008) with an intensity increase of 5.2 and 5.7 in comparison to the control. c) Unknown B at 944.360964 *m/z* and 311 s ($P_{CNV}$=0.005; $P_{PCV}$=0.03) with an intensity increase of 1.7 and 1.9 respectively. d) Significant deregulated compound 5,8,11,14-Eicosatetraenoic acid (EPA - $C_{20}H_{32}O_2$ - based on putative identification) ($P_{CNV}$=0.04; $P_{PCV}$=0.01) with an increase in mean intensity of 1.4 and 1.7 times in contrast to the control. d) Significant deregulated compound 4,7,10,13,16,19-Docosahexaenoic acid (DHA - $C_{22}H_{32}O_2$ - based on putative identification) ($P_{CNV}$=0.008; $P_{PCV}$=0.006) with an increase mean intensity of 1.7 and 2.0 times in contrast to the control. The identification of the compounds is based on MS1 accurate mass search and MS2 fragment annotation.

The identification results of the differential expression analysis are based on putative identifications via MS1 accurate mass search and MS2 fragment annotation, corresponding to a level 3 identification[84]. Here, to reach a level 1 identification, additional experiments in follow up studies are necessary to validate the potential biomarkers.

### 3.3.6 Limitations and Runtime of DIAMetAlyzer

It can be deemed a limitation that DDA and DIA data has to be measured for an experiment. The main purpose of the DIAMetAlyzer workflow is to perform accurate quantification in a

targeted manner. Here, the DDA data - for example - reference standards would be measured once to construct the assay library. This library can then be reused for DIA data analysis measured with the same experimental setup. In a targeted setting, it is generally necessary to invest resources to build accurate assays in order to achieve high-quality targeted results. While DDA is generally biased towards high abundant analytes, this will not impact measurements of low complexity, such as pure standards. When building assay libraries from complex samples, the library will be biased towards highly abundant analytes. We suggest to counteract this bias by enhancing such assay libraries with reference compounds measured from pure standards. DIAMetAlyzer uses SIRIUS for fragment annotation, so the limitations in terms of high-resolution instruments and molecular masses of SIRIUS apply to the workflow as well. High mass compounds can in some cases not be processed by SIRIUS in a timely manner. The user can set a threshold of 100 s (default) per compound, to restrict the runtime. As a reference, the assay library generated from 67 DDA samples, with prior MS1 identification took around 2.5 h using 10 cores (Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz). With allowing unknown features it took around 12.5 h using 28 cores. The runtime of the complete KNIME workflow for the targeted pesticide mix experiment, using one core (Intel Core i7 @ 3.50 GHz), was 36 minutes. All runtime improvements of SIRIUS in the future will also impact the runtime of the workflow. With the integration of the *AssayGeneratorMetabo* into OpenMS, we provide an easy-to-use solution for target-decoy assay library generation in OpenMS using the fragmentation re-rooting method[1]. However, combining multiple feature detection methods as we did for MetaboDIA and the DIAMetAlyzer is not straightforward due to interoperability issues between the tools. For this purpose, we provide means to add decoys on the assay library level (*DecoyGeneratorMetaboTool*). For further details regarding the decoy methods on library level please see Supplementary Fig. A.8. The so generated target and decoy assay library can then be appended to the one from the *AssayGeneratorMetabo*. The combined library can be used in the DIAMetAlyzer workflow for the DIA data analysis.

## 3.4 Discussion

In conclusion, we present a novel analysis workflow for metabolomics DIA data that introduces accurate control of the FDR for the first time. Our workflow is based on industry-grade computational libraries and workflow engines (OpenMS and KNIME) and builds on existing open-source software. It is possible to use the OpenMS command-line tools and algorithms to build the workflow in any scripting environment, cluster or cloud infrastructure. Our adaptive machine learning approach integrates the signal from MS1 and MS2 levels to optimally separate true signal from noise and provide a well-calibrated FDR estimate. In the past, DIA data was complicated to analyze for mass spectrometry practitioners and available tools could result in vastly different results due to the lack of reliable estimates of precision in the reported data.

Introducing a standardized workflow that utilizes a statistically well-calibrated FDR will allow practitioners to analyze and compare DIA data on equal footing. This novel extension allows for improvements in the reliability and robustness of metabolomics discovery. Importantly, our pipeline can be used in a targeted setting (quantifying known compounds) as well as in an untargeted setting (quantifying unknown compounds using their m/z patterns). In comparison to MS-DIAL, a software for untargeted deconvolution, we were able to detect almost twice as many compounds in the targeted setting. In comparison to MetaboDIA, a tool for consensus spectral library building for metabolomics data from DDA data, we could quantify 110 additional features while combining the two libraries, further increasing the number of quantified features by 132. Extracting both known and unknown features (without prior identification), we further improved the number of quantified features by 32%. As shown in the analysis of data from AMD patients, this can lead to new biological findings using an unknown library. Specifically, using an experimentally specific DDA library based on reference substances allows for the accurate identification of compounds and markers from DIA data in low concentrations, facilitating biomarker quantification.

# Chapter 4

# Reporting Standardization in Metabolomics: MzTab-M

This chapter includes partially identical or adapted content with permission from:

---

*mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics*

Nils Hoffmann, Joel Rein, Timo Sachsenberg, Jürgen Hartler, Kenneth Haug, Gerhard Mayer, Oliver Alka, Saravanan Dayalan, Jake T. M. Pearce, Philippe Rocca-Serra, Da Qi, Martin Eisenacher, Yasset Perez-Riverol, Juan Antonio Vizcaino, Reza M. Salek, Steffen Neumann, and Andrew R. Jones

*Analytical Chemistry* 91 (5), 3302-3310 (2019)

---

A detailed description of the contributions to the project by coauthors is provided in the Appendix C

## 4.1 Introduction

In the last decades, the exchange of scientific data within the proteomics and metabolomics community was aggravated due to inconsistency in data exchange formats. The Human Proteome Organization (HUPO) initiated a standardization endeavour to establish standardized exchange formats for different stages along the data processing pipeline by introducing the Proteomics Standards Initiative (PSI). To this date, they provide standardized XML and table-based data formats for the MS field, for example, for raw data (mzML), identification data (mzIdentML), and quantification data (mzQuantML) as well as the table-based analysis summary output format called *MzTab*. This effort goes hand in hand with the recently introduced

FAIR principles, aiming for the data of a study to be findable, accessible, interoperable, and reusable[3,4]. Allowing such reusability of data needs the formulation of non-proprietary agreed standard formats. With the introduction of the standardized reporting format *MzTab* in 2014, the reporting and exchange of proteomics and metabolomics research was simplified[5]. Over the years it grew apparent that the current *MzTab* standard was not covering the reporting necessities of the fast-growing metabolomics community[6]. This was mainly due to the heterogeneity of metabolomics identification which could not be described in *MzTab*. To facilitate a more comprehensive exchange, the standard format *MzTab-M* was introduced for metabolomics data as a joint endeavour of the Metabolomics Standard Initiative (MSI) and the PSI. Here, we present an overview of the new reporting standard for metabolomics and its implementation in the OpenMS framework.

## 4.2 Methods

### 4.2.1 Rationals

The development of *MzTab-M* was a community-driven effort guided by the following rationals. The data format is intended to:

- facilitate accessibility and data sharing.
- contain sufficient information to be used as standard documentation format for supplementary-material sections of publications in metabolomics.
- report on different levels from experimental design over quantification and identification to a summary.
- be inspected with software such as Microsoft Excel or Open Office Spreadsheet.
- be an output format for web-services readily accessed by tools supporting *MzTab-M*.
- directly link a small molecule record to its spectrum in an external MS data file.
- have results easily accessible to scripting languages allowing bioinformaticians to develop software.
- contain the complete final results of an MS-based metabolomics experiment in a single file.

### 4.2.2 Structure

*MzTab-M* is structured in four cross-referenced tables (see Fig. 4.1, 4.3):

- **MTD** Metadata Table
- **SML** Small Molecule Table
- **SMF** Small Molecule Feature Table
- **SME** Small Molecule Evidence Table

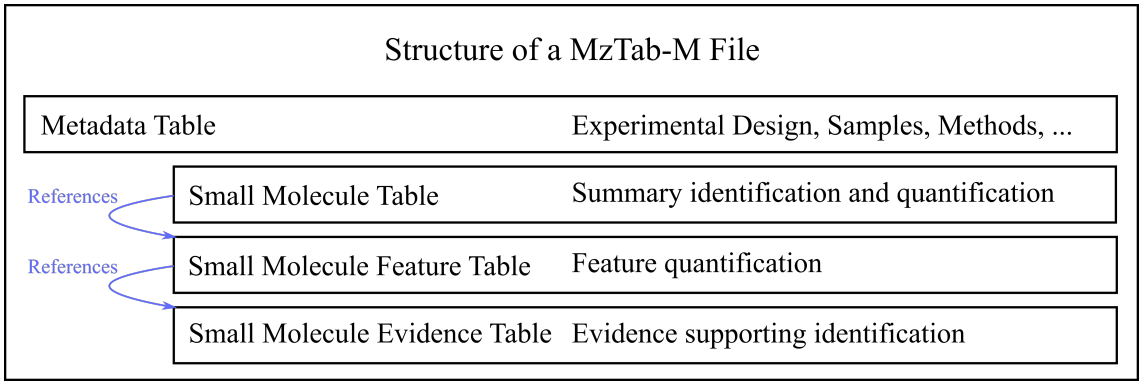


**Figure 4.1: *MzTab-M* Structure** The Metadata Table contains all metadata of the experiment and its analysis, including the experimental design. The Small Molecule Table is a summary of all identification and quantification results. The Small Molecule Feature Table holds the information of all quantified features. The Small Molecule Evidence Table stores identification data and the evidence supporting the assigned identifications. The table entries of different levels are referenced from the highest (Summary) to the lowest level (Evidence), to allow for multiple quantifications and heterogenous identifications.

### 4.2.3 Metadata Table

The MTD table stores metadata for tracing the experimental steps and the data analysis. This is represented in three columns, the first one is the identifier "MTD" column, the second contains a parameter name and the third the corresponding parameter value (Table 4.1). Most of the mandatory entries are represented by controlled vocabulary (CV), which is a standardized specification of e.g., the instrument type used or the polarity of the experiment. In addition, non-mandatory entries, such as the author of the file, can be provided. The most important entry is the storage of the experimental design which can help with further data analysis and visualization.

The basic structure of such as design is described by the following aspects:

- **Sample:** Analyzed biological material. It can be described, among others, as one or multiple species, cells, and tissue samples.
- **MS_run:** Represents a single run on an MS instrument identified by file name and format.

- **Assay:** A measurement of a sample that produced quantitative values of small molecules or lipids. In the case of fractionated experiments, an assay can represent multiple MS runs/samples.
- **Study_variable:** Represents an experimental condition e.g., case, control. Depending on the experimental design, it can represent a group of replicates which may be averaged from individual assays.

A common example for an experimental design would be a study in which two conditions are compared, e.g., case and control. The structure of such a design is shown in Fig. 4.2. Here, two groups, study_variable 1 (e.g., case) and study_variable 2 (e.g., control) are compared by measuring three biological replicates each (assay 1-3/sample 1-3, assay 4-6/sample 4-6), but no technical replicates. Each replicate analyzed by MS is represented by an ms_run (ms_run 1-6).

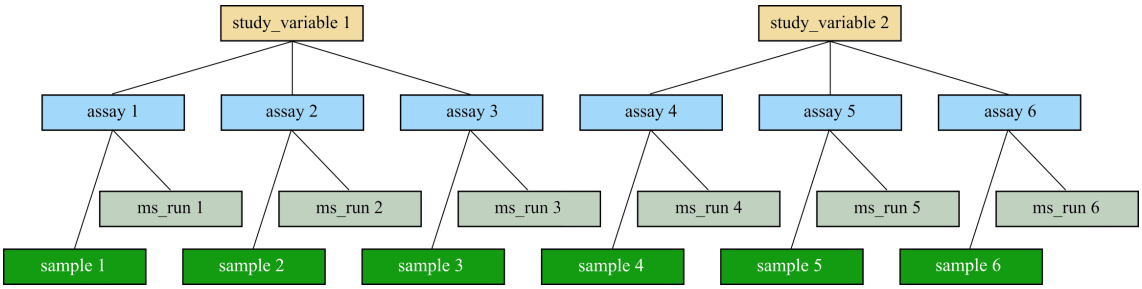


**Figure 4.2: Experimental Design** Depicts two conditions: study_variable 1 (e.g., case) and study_variable 2 (e.g., control). With three biological replicates each (assay 1-3/sample 1-3, assay 4-6/sample 4-6), measured over 6 runs on the MS (ms_run 1-6).

In addition to the experimental design, optional columns can be added to the MTD table describing metadata useful for study interpretation, such as sample processing steps, software, parameters and contact details. Other (specialized) formats containing the experimental design and sample processing information might be used instead, such as ISA-TAB, and referenced in the MTD section.

**Table 4.1: Metadata Table**

| | | |
|---|---|---|
| MTD | mzTab-version | 2.0.0-M |
| MTD | mzTab-ID | local_id: 14128832103716313886 |
| MTD | software[1] | [MS, MS:1001456, analysis software, 2.6.0-pre-idf-ams-2021-12-09] |
| MTD | software[2] | [MS, MS:1002169, TOPP FeatureFinderMetabo, 2.4.0-nightly-2019-07-17] |
| MTD | software[3] | [MS, MS:1001456, analysis software, 2.4.0-nightly-2019-07-17] |
| MTD | quantification_method | [MS, MS:1001834, LC-MS label-free quantitation analysis, ] |
| MTD | ms_run[1]-location | file://PestMix1_1ngSolventDDA20-50.wiff |
| MTD | ms_run[1]-scan_polarity[1] | [MS, MS:1000130, positive scan, ] |
| MTD | assay[1] | assay_PestMix1_1ngSolventDDA20-50 |
| MTD | assay[1]-ms_run_ref | ms_run[1] |
| MTD | study_variable[1] | study_variable_PestMix1_1ngSolventDDA20-50 |
| MTD | study_variable[1]-assay_refs | assay[1] |
| MTD | study_variable[1]-description | study_variable_PestMix1_1ngSolventDDA20-50 |
| MTD | cv[1]-label | PSI-MS |
| MTD | cv[1]-full_name | MS |
| MTD | cv[1]-version | 4.1.49 |
| MTD | cv[1]-uri | https://raw.githubusercontent.com/HUPO-PSI/psi-ms-CV/master/psi-ms.obo |
| MTD | database[1] | [, , CustomDB, ] |
| MTD | database[1]-prefix | CustomDB |
| MTD | database[1]-version | 0 |
| MTD | database[1]-uri | file:///Users/alka/Desktop/AMS_ID_test/AMS_test_Mapping.tsv |
| MTD | small_molecule-quantification_unit | [MS, MS:1001844, MS1 feature area, ] |
| MTD | small_molecule_feature-quantification_unit | [MS, MS:1001844, MS1 feature area, ] |
| MTD | small_molecule-identification_reliability | [MS, MS:1002896, compound identification confidence level, ] |
| MTD | id_confidence_measure[1] | [, , MassErrorPPMScore, ] |
| MTD | id_confidence_measure[2] | [, , MassErrorDaScore, ] |

### 4.2.4 Small Molecule Table

The SML table is a summary of the quantification and - if available - identification results (Table 4.2). The first line of the table is the Small Molecule Header (SMH) defining the columns. Afterwards, the SML entries are added row by row, with the first cell containing 'SML'. Each row starts with a unique identifier for each analyte, the SML_ID. It is followed by the SMF_ID_REFS referencing the corresponding features in the SMF table, which derives from the same molecule with different adduct forms or in-source fragments. The subsequent columns represent the identification of the compound such as database_identifier, representing the identifier of the compound in the database used for identification, the chemical_formula presenting the molecular formula of the compound, smiles providing the SMILES[85] representation, inchi adding the InChi[86] representation, chemical_name representing the name of the identified analyte, uri, if possible, linking to the website of the database entry, and theoretical_neutral_mass providing the theoretical neutral mass of the compound. The consecutive columns report the quantitative results, where each assay (abundance_assay) and each study variable (abundance_study_variable e.g., as mean across the assay values within a study_variable) is represented in a column. In addition, the variability of the study variable (e.g., the standard error) can be reported (abundance_variation_study_variable). The unit and interpretation of the quantitative values are represented in the MTD. Furthermore, user-specified optional columns can be added to improve the flexibility of the format.

**Table 4.2: Small Molecule Table**

| **SMH** | **SML_ID** | **SMF_ID_REFS** | **database_identifier** | **chemical_formula** | **smiles** | **...** |
|---|---|---|---|---|---|---|
| SML | 1 | 1 | 99675-03-3 | C14H22N1O4P1S1 | CC(C)NP.. | ... |
| **...** | **inchi** | **chemical_name** | **uri** | **theoretical_neutral_mass** | **adduct_ions** | **...** |
| ... | InChI=1S/C14H22NO4PS/c1... | Isofenphos-methyl | null | 331.100717 | [M+H]1+ | ... |
| **...** | **reliability** | **best_id_confidence_measure** | **best_id_confidence_value** | **abundance_assay[1]** | | **...** |
| ... | 2 | null | null | 7.71E+04 | | ... |
| **...** | **abundance_study_variable[1]** | **abundance_variation_study_variable[1]** | | | | |
| ... | null | null | | | | |

**Table 4.3: Small Molecule Feature Table**

| **SFH** | **SMF_ID** | **SME_ID_REFS** | **SME_ID_REF_ambiguity_code** | **adduct_ion** | **isotopomer** | **...** |
|---|---|---|---|---|---|---|
| SMF | 1 | 1 | null | [M+H]1+ | null | ... |
| **...** | **exp_mass_to_charge** | **charge** | **retention_time_in_seconds** | **retention_time_in_seconds_start** | **retention_time_in_seconds_end** | **...** |
| ... | 332.108359 | 1 | 518.841 | null | null | ... |
| **...** | **abundance_assay[1]** | **opt_global_FWHM** | | | | |
| ... | 7.71E+04 | 2.57700372 | | | | |

### 4.2.5 Small Molecule Feature Table

The SMF table consists of information about the features measured by the instrument and quantified by the software (Table 4.3). The first row, the Small Molecule Feature Header (SFH), explains the column entries. The SMF entries follow afterwards in a row-wise manner. Each row represents a feature entry, which was linked over multiple MS runs, with missing values handled accordingly (i.e., by representing them as *null*). The specification document also describes how non-aligned workflows can be represented. The SME_ID_REFS refer to the Small Molecule Evidence (SME), which is the identification evidence linked to a specific aligned feature. In the case of ambiguous identification, an additional code (SME_ID_REF_ambiguity_code) depicts how the reader should handle this case. Further entries in the SMF table cover information about the adduct (adduct), charge (charge), experimental m/z value (exp_mass_to_charge), retention time (retention_time_in_seconds, retention_time_in_seconds_start, retention_time_in_seconds_end) and additional information about the quantified peak (isotopomer), followed by the quantitative values for each assay specified (abundance_assay) in the MTD. Again, additional optional columns can be used to add further information to the table.

### 4.2.6 Small Molecule Evidence Table

The SME table depicts potentially ambiguous types of evidence supporting the identification of a certain molecule (Table 4.4). Each row consists of the results of one identification process, such as spectral library search, *de-novo* identification, accurate mass search, and manual curation. The table header is represented by the small molecule evidence header (SEH) and the columns describe the entry types. The SME columns consist of a local identifier (SME_ID), followed by entries for the input data (evidence_input_id). This is needed if multiple rows are reported for the same input data based on the MS2 spectrum, retention time, and accurate mass search results. They can be linked by the same entry, referred to as evidence_input_id. Similar to the SML table, the columns represent the identification of the analyte based on the database identifier (database_identifier), chemical formula (chemical_formula ), SMILES[85] (smiles), InChi[86] (inchi), chemical name (chemical_name), and uri (uri). In addition, the experimental m/z (exp_mass_to_charge), the charge (charge), and the theoretical m/z (theoretical_mass_to_charge) can be recorded together with the scores, or confidence measures (id_confidence_measure) by the software used for the identification (identification_method). The exact spectrum used for the identification can be represented via the source file and its index or native id (spectra_ref).

**Table 4.4: Small Molecule Evidence Table**

| **SEH** | **SME_ID** | **evidence_input_id** | **database_identifier** | **chemical_formula** | **...** |
|---|---|---|---|---|---|
| SME | 1 | mass=332.1083548,rt=518.8410008 | 99675-03-3 | C14H22N1O4P1S1 | ... |
| ... | **smiles** | **inchi** | **chemical_name** | **uri** | **...** |
| ... | CC(C)NP... | InChI=1S/C14H22NO4PS/c1... | Isofenphos-methyl | null | ... |
| **...** | **derivatized_form** | **adduct_ion** | **exp_mass_to_charge** | **charge** | **...** |
| ... | null | [M+H]1+ | 332.108359 | 1 | ... |
| **...** | **theoretical_mass_to_charge** | **spectra_ref** | **identification_method** | **ms_level** | **...** |
| ... | 331.100717 | ms_run[1]:248283675846367528 | [MS, MS:1000207, accurate mass, ] | [MS, MS:1000511, ms level, 1] | ... |
| **...** | **id_confidence_measure[1]** | **id_confidence_measure[2]** | **rank** | | |
| ... | 1 | 1 | 1 | | |

### 4.2.7 Identification Evidence and Ambiguity

Correct identification of an analyte in metabolomics is still a challenge. Identification can be based on different kinds of evidence, for example, accurate mass, or the MS2 spectrum via spectral library search. The *MzTab-M* format accommodates all different possibilities in a flexible structure based on the SME level. In the final report, the export software can include one or more database identifiers, which are referenced in the MTD. In addition, the molecular formula is stored in standard notation and via simplified molecular-input line-entry system SMILES[85] or InChi[86]. If the ambiguity of the identification cannot be resolved, it can be represented by a pipe-separated list of identifiers. Several measures are available for describing the confidence of an identification. Here, the reliability codes developed by the MSI[84,87] and the scores or confidence measures from the identification software can be used. The evidence source can be traced via references to the SMF table. If adduct grouping was performed previously, the same SME row may point to multiple SMF instances. If multiple identification procedures were performed for the same feature, it is expected that the SMF element will reference multiple SME elements while sharing the evidence_input_id. In this case, multiple SME identifiers are referenced by one SMF an additional code (SME_ID_REF_ambiguity_code) can be provided, indicating whether there is an ambiguity in the identifications (Fig. 4.3).

| SMH | SML_ID | SMF_ID_REFS | database_identifier | theoretical_neutral_mass | abundance_assay[1] |
|---|---|---|---|---|---|
| SML | 1 | 1 | HMDB:HMDB00043\|HMDB:HMDB00883\|HMDB:HMDB01382 | 117.0789794 | 406.5467529 |
| SML | 2 | 2 | HMDB:HMDB00162\|HMDB:HMDB03411\|HMDB:HMDB12880 | 115.0633293 | 227.974411 |
| SML | 3 | 3 | HMDB:HMDB05033 | 314.0725142 | 464.2641296 |
| SML | 4 | 4 | HMDB:HMDB30800 | 196.0371753 | 172.7888947 |
| SML | 5 | 5 | HMDB:HMDB41777\|HMDB:HMDB41778 | 380.0202061 | 464.3032227 |

| SFH | SMF_ID | SME_ID_REFS | SME_ID_REF_ambiguity_code | exp_mass_to_charge | retention_time_in_seconds | abundance_assay[1] |
|---|---|---|---|---|---|---|
| SMF | 1 | 1\|2\|3 | 1 | 118.0862817 | 70.1570034 | 406.5467529 |
| SMF | 2 | 4\|5\|6 | 1 | 116.0705021 | 71.35399818 | 227.974411 |
| SMF | 3 | 7 | null | 337.0614392 | 72.79500246 | 464.2641296 |
| SMF | 4 | 8 | null | 219.026404 | 76.87600136 | 172.7888947 |
| SMF | 5 | 9\|10 | 1 | 403.0090757 | 77.83500195 | 464.3032227 |

| SEH | SME_ID | evidence_input_id | database_identifier |
|---|---|---|---|
| SME | 1 | mass=118.086,rt=70.157 | HMDB:HMDB00043 |
| SME | 2 | mass=118.086,rt=70.157 | HMDB:HMDB00883 |
| SME | 3 | mass=118.086,rt=70.157 | HMDB:HMDB01382 |
| SME | 4 | mass=116.070,rt=71.353 | HMDB:HMDB00162 |
| SME | 5 | mass=116.070,rt=71.353 | HMDB:HMDB03411 |
| SME | 6 | mass=116.070,rt=71.353 | HMDB:HMDB12880 |
| SME | 7 | mass=337.061,rt=72.795 | HMDB:HMDB05033 |
| SME | 8 | mass=219.026,rt=76.876 | HMDB:HMDB30800 |
| SME | 9 | mass=403.009,rt=77.835 | HMDB:HMDB41777 |
| SME | 10 | mass=403.009,rt=77.835 | HMDB:HMDB41778 |

**Figure 4.3: Example of *MzTab-M* Referencing and Identification Ambiguity Representation** The referencing between the SML and SMF table uses the SMF_ID_REFS. In this example one file was analysed, leading to a 1:1 mapping of the references in the summary table. SME_ID_REFS are used to reference quantiative feature information (SMF) with their corresponding identification (SME). In this example, for the first feature three ambiguous identifications were assessed using accurate mass search, leading to the reference of three SME_IDs to one SMF_ID.

### 4.2.8 Controlled Vocabulary and File Validation

CV terms are used to provide unambiguous annotations. PSI-MS is used for terms associated with MS and processing[88]. We have extended the PSI semantic validation framework, to ensure the validity of the CV terms[89]. A mapping file is included in the framework, stating groups of CV terms allowed at certain positions in the *MzTab-M* File. New terms can be added via a pull request or request to the mailing list. To ensure the validity of the *MzTab-M* file and the CV terms used by the software, a validation tool jmzTab (project: https://github.com/lifs-tools/jmzTab-m) was developed.

### 4.2.9 Implementation in Software and Databases

The stable version (Mztab-M 2.0) has been released after a review process from PSI and MSI. Further changes to the format are not expected in the following years. A reference implementation with parser, writer, and validator (in jmzTab-m) has been developed in Java. jmzTab-m provides an OpenAPI 2.0 compatible API model that serves as the basis for automatic model generation in a wide number of programming languages (C++, JavaScript, R, Python), reducing the burden of implementation. A user-friendly web-based application (https://apps.lifs.isas.de/mztabvalidator/) was provided to allow easy semantic validation and control. Additional implementations are under development in different software frameworks, including XCMS[90], Lipid Data Analyzer[91], OpenMS[7] and MetaboLights[92]. Over the coming years, we will be promoting the implementation of the standard in a wide variety of both open-source and commercial software to act as a universal standard for metabolomics and lipidomics.

### 4.2.10 Implementation in OpenMS

The implementation has to fulfill the requirements concerning the semantic representation of the data and the data format specified in the *Mztab-M* specification (`https://github.com/HUPO-PSI/mzTab/blob/master/specification_document-releases/2_0-Metabolomics-Release/mzTab_format_specification_2_0-M_release.adoc`). In the OpenMS implementation the data model (in-memory representation) is separated from the serialization into a *Mztab-M* textfile. OpenMS already implements the previous standard *MzTab* (`https://github.com/HUPO-PSI/mzTab/blob/master/specification_document-releases/1_0-Proteomics-Release/mzTab_format_specification.pdf`). Existing code, for functionalities used by *MzTab* and *MzTab-M*, was moved to a base class reducing code duplication.

### *MzTab-M* data model

***MzTabM*** The *MzTab-M* data model was implemented based on the format specifications provided by the *MzTab-M* standard documentation (`https://github.com/HUPO-PSI/mzTab/blob/master/specification_document-releases/2_0-Metabolomics-Release/mzTab_format_specification_2_0-M_release.adoc#format-specification`). The class *MzTabM* unites all relevant information in its data model, to allow the storage independent of the data format used for serialization (Fig. 4.4). It inherits from the *MzTabBase* as a common denominator of the *MzTab* and the *MzTabM* implementation. The *MzTabBase* class has one protected member function (*MzTabBase::getOptionalColumnNames_()*). This function is used in the *MzTab* and the *MzTabM* class providing a common access function for optional column names. The member variables of *MzTabM* represent the standardized *MzTab-M* sections (metadata section (*MzTabMMetaData*), small molecule section (*MzTabMSmallMoleculeSectionRow*), small molecule feature section (*MzTabMSmallMoleculeFeatureSectionRow*), and small molecule evidence section (*MzTabMSmallMoleculeEvidenceSectionRow*), which are implemented as individual classes used by *MzTabM*. Additionally, members for empty rows, comments, and optional column names exist. Various getters and setters are provided to give access to protected member variables. The function *MzTabM::exportFeatureMapToMzTabM()* allows the export of data held in an *OpenMS::FeatureMap* into the *MzTab-M* data model. Additional protected member functions were added. The first to get an adduct as a string representation from the internal identification data (*MzTabM::getAdductString_()*). The second to provide the ability to access and store associated meta values from an *OpenMS::FeatureMap*, which stores quantitative and identification information in optional columns (adding the value and the optional column name to the appropriate section)(*MzTabM::getFeatureMapMetaValue()*).

MzTabBase
MzTabBase()
virtual ~MzTabBase()
std::vector<String> getOptionalColumnNames_()
MzTabM
MzTabMMetaData m_meta_data_
MzTabMSmallMoleculeSectionRows m_small_molecule_data_
MzTabMSmallMoleculeFeatureSectionRows m_small_molecule_feature_data_
MzTabMSmallMoleculeEvidenceSectionRows m_small_molecule_evidence_data_
std::vector<Size> empty_rows_
std::map<Size, String> comment_rows_
std::vector<String> sml_optional_column_names_
std::vector<String> smf_optional_column_names_
std::vector<String> sme_optional_column_names_
MzTabM()
~MzTabM()
const MzTabMetaData& getMetaData()
void setMetaData()
const MzTabMSmallMoleculeSectionRows& getMSmallMoleculeSectionRows()
void setMSmallMoleculeSectionRows()
const MzTabMSmallMoleculeFeatureSectionRows& getMSmallMoleculeFeatureSectionRows()
void setMSmallMoleculeFeatureSectionRows()
const MzTabMSmallMoleculeEvidenceSectionRows& getMSmallMoleculeEvidenceSectionRows()
void setMSmallMoleculeEvidenceSectionRows()
void setCommentRows()
void setEmptyRows()
const std::vector<Size>& getEmptyRows()
const std::map<Size, String>& getCommentRows()
std::vector<String> getMSmallMoleculeOptionalColumnNames()
std::vector<String> getMSmallMoleculeFeatureOptionalColumnNames()
std::vector<String> getMSmallMoleculeEvidenceOptionalColumnNames()
void addMetaInfoToOptionalColums()
MzTabM exportFeatureMapToMzTabM()
String getAdductString_()
void getFeatureMapMetaValues_()
MzTabMMetaData
MzTabString mz_tab_version
MzTabString mz_tab_id
MzTabString title
MzTabString description
std::vector<MzTabParameter> sample_processing
std::map<Size, MzTabInstrumentMetaData> instrument
std::map<Size, MzTabSoftwareMetaData> software
std::map<Size, MzTabString> publication
std::map<Size, MzTabContactMetaData> contact
std::map<Size, MzTabString> uri
std::map<Size, MzTabString> external_study_uri
MzTabParameter quantification_method
std::map<Size, MzTabSampleMetaData> sample
std::map<Size, MzTabMMSRunMetaData> ms_run
std::map<Size, MzTabMAssayMetaData> assay
std::map<Size, MzTabMStudyVariableMetaData> study_variable
std::map<Size, MzTabParameter> custom
std::map<Size, MzTabDoubleList> cv
std::map<Size, MzTabMDatabaseMetaData> database
std::map<Size, MzTabParameter> derivatization_agent
MzTabParameter small_molecule_quantification_unit
MzTabParameter small_molecule_feature_quantification_unit
MzTabParameter small_molecule_identification_reliability
std::map<Size, MzTabParameter> id_confidence_measure
std::vector<MzTabString> colunit_small_molecule
std::vector<MzTabString> colunit_small_molecule_feature
std::vector<MzTabString> colunit_small_molecule_evidence
MzTabMSmallMoleculeSectionRow
MzTabString sml_identifier
MzTabStringList smf_id_refs
MzTabStringList database_identifier
MzTabStringList chemical_formula
MzTabStringList smiles
MzTabStringList inchi
MzTabStringList chemical_name
MzTabStringList uri
MzTabDoubleList theoretical_neutral_mass
MzTabStringList adducts
MzTabString reliability
MzTabParameter best_id_confidence_measure
MzTabDouble best_id_confidence_value
std::map<Size, MzTabDouble> small_molecule_abundance_assay
std::map<Size, MzTabDouble> small_molecule_abundance_study_variable
std::map<Size, MzTabDouble> small_molecule_abundance_variation_study_variable
std::vector<MzTabOptionalColumnEntry> opt_
MzTabMSmallMoleculeFeatureSectionRow
MzTabString smf_identifier
MzTabStringList sme_id_refs
MzTabInteger sme_id_ref_ambiguity_code
MzTabString adduct_ion
MzTabParameter isotopomer
MzTabDouble exp_mass_to_charge
MzTabInteger charge
MzTabDouble retention_time
MzTabDouble rt_start
MzTabDouble rt_end
std::map<Size, MzTabDouble> small_molecule_feature_abundance_assay
std::vector<MzTabOptionalColumnEntry> opt_
MzTabMSmallMoleculeEvidenceSectionRow
MzTabString sme_identifier
MzTabString evidence_input_id
MzTabString database_identifier
MzTabString chemical_formula
MzTabString smiles
MzTabString inchi
MzTabString chemical_name
MzTabString uri
MzTabParameter derivatized_form
MzTabString adduct
MzTabDouble exp_mass_to_charge
MzTabInteger charge
MzTabDouble calc_mass_to_charge
MzTabSpectraRef spectra_ref
MzTabParameter identification_method
MzTabParameter ms_level
std::map<Size, MzTabDouble> id_confidence_measure
MzTabInteger rank
std::vector<MzTabOptionalColumnEntry> opt_

**Figure 4.4: Class Diagram of *MzTabM*** *MzTabM* inherits from *MzTabBase*, which is the base for the *MzTab* and *MzTabM* implementation. It provides one protected member function (*MzTabBase::getOptionalColumnNames_()*) for a common access function of optional column names. The member variables of *MzTabM* represent the standardized *MzTab-M* sections (metadata section, small molecule section, small molecule feature section, small molecule evidence section), which are implemented as individual classes used by *MzTabM*. Additionally, members for empty rows, comments, and optional column names exist. Various getters and setters are provided to give access to protected member variables. The function *MzTabM::exportFeatureMapToMzTabM()* allows the export of data held in an *OpenMS::FeatureMap* into the *MzTab-M* data model. *MzTabM::getAdductString_()* get an adduct as a string representation. *MzTabM::getFeatureMapMetaValue()* provides the ability to access and store associated meta values from an *OpenMS::FeatureMap* in optional columns (adding the value and the optional column name to the appropriate section).

A small example is provided to show how data stored in an OMSFile (.oms), which is an SQLite-based storage container that can store feature and identification information, can be exported to the MzTab-M data model (Listing 4.1).

**Listing 4.1:** Code Snippet Presenting the Usage of *MzTabM*

```
1  FeatureMap featuremap;
2  MzTabM mztabm;
3
4  // load the data stored in an OMSFile (SQLite-based format used for identification
       data) to a FeatureMap.
5  OMSFile().load(OPENMS_GET_TEST_DATA_PATH("MzTabMFile_input_1.oms"), featuremap);
6
7  // export a FeatureMap with IdentificationData to MzTabM (data model)
8  mztabm = mztabm.exportFeatureMapToMzTabM(featuremap);
```

***MzTabMetaData*** The *MzTabMetaData* class stores additional information about the data set. A detailed description of the information to store can be found here `https://github.com/HUPO-PSI/mzTab/blob/master/specification_document-releases/2_0-Metabolomics-Release/mzTab_format_specification_2_0-M_release.adoc#62-metadata-section`. The member variables of *MzTabMetaData* represent the standardized *MzTab-M* metadata section (Fig. 4.5). The information can either be stored in basic MzTab datatypes (e.g., *MzTabString*, *MzTabInteger*, *MzTabParameter*), their respective list types (e.g., *MzTabParameterList*) or more complex datatypes using the basic ones (e.g., *MzTabInstrumentMetaData*, *MzTabSoftwareMetaData*, *MzTabContactMetaData*, *MzTabSampleMetaData*, *MzTabCVMetaData*, *MzTabMMSRunMetaData*, *MzTabMAssayMetaData*, *MzTabMStudyVariableMetaData*, *MzTabMDatabaseMetaData*). All basic datatypes and some complex ones were implemented with the introduction of *MzTab* to OpenMS others were added to allow the implementation of the *MzTab-*

*M* standard (e.g., classes starting with *MzTabM*). The basic types are used in all complex classes of (*MzTab* and *MzTabM*) either alone or in combination with various C++ standard types (e.g., *std::map*), depending on the information stored in the datatype/class. The MzTab basic types provide getters and setters to the protected member variables, as well as additional checks (e.g., *::isNull()*, *::isNaN()*, *::isInf()*) where appropriate. Additionally, they provide the functions *toCellString()* and *fromCellString()*, allowing the transfer between the basic type and its string representation.

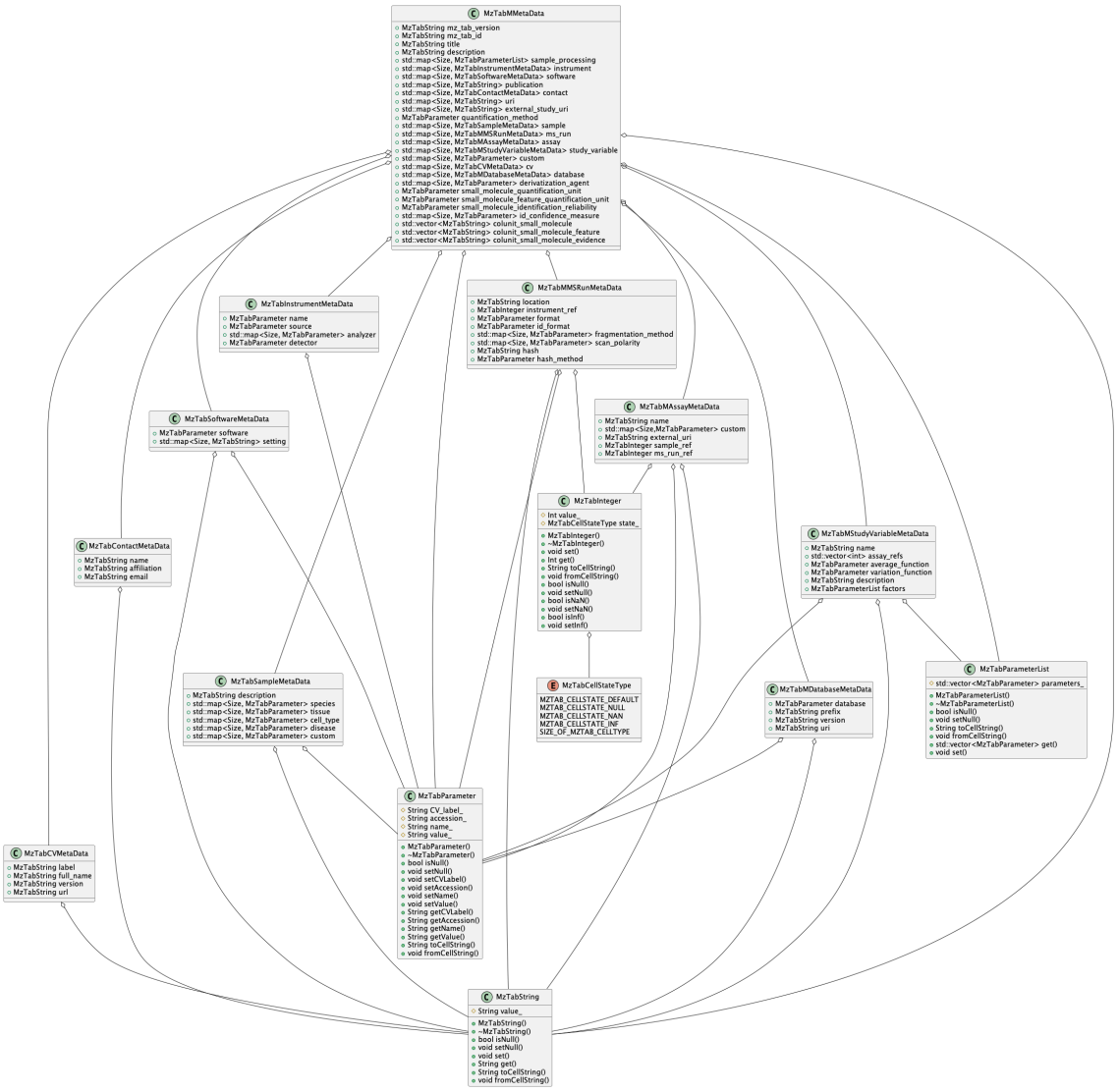


**Figure 4.5: Class Diagram of *MzTabMMetaData*** The member variables of *MzTabMetaData* represent the standardized *MzTab-M* metadata section. The information can either be stored in basic MzTab datatypes (e.g., *MzTabString*, *MzTabInteger*, *MzTabParameter*), their respective list types (e.g., *MzTabParameterList*) or more complex datatypes using the basic ones (e.g., *MzTabInstrumentMetaData*, *MzTabSoftwareMetaData*, *MzTabContactMetaData*, *MzTabSampleMetaData*, *MzTabCVMetaData*, *MzTabMMSRunMetaData*, *MzTabMAssayMetaData*, *MzTabMStudyVariableMetaData*, *MzTabMDatabaseMetaData*). The MzTab basic types provide getters and setters to the protected member variables, as well as additional checks (e.g., *::isNull()*, *::isNaN()*, *::isInf()*) were appropriate. Additionally, they provide the functions *toCellString()* and *fromCellString()*, allowing the transfer between the basic type and its string representation.

A small example is provided to show how data can be added to the *MzTabMetaData* and the *MzTabM* instance (Listing 4.2).

**Listing 4.2:** Code Snippet Presenting the Usage of *MzTabMMetaData*

```
MzTabM mztabm;

// metadata
MzTabMMetaData mztabm_meta;
mztabm_meta.mz_tab_id.set("local_identifier");
mztabm_meta.title.set("SML_ROW_TEST");
mztabm_meta.description.set("small_molecule_section_row_test");

// instrument
MzTabInstrumentMetaData meta_instrument;
meta_instrument.name.fromCellString("[MS, MS:1000483, Thermo Fisher Scientific
    instrument model, LTQ Orbitrap Velos]");
meta_instrument.source.fromCellString("[MS, MS:1000008, Ionization Type, ESI]");
MzTabParameter ana;
ana.fromCellString("[MS, MS:1000443, Mass Analyzer Type, Orbitrap]");
meta_instrument.analyzer[0] = ana;
meta_instrument.detector.fromCellString("[MS, MS:1000453, Detector, Dynode
    Detector]");
mztabm_meta.instrument[0] = meta_instrument;

// software
MzTabSoftwareMetaData meta_software;
MzTabParameter p_software;
p_software.fromCellString("[MS, MS:1002205, ProteoWizard msconvert, ]");
meta_software.software = p_software;
meta_software.setting[0] = MzTabString("Peak Picking MS1");
mztabm_meta.software[0] = meta_software;

// sample
MzTabSampleMetaData meta_sample;
meta_sample.description = MzTabString("Nice Sample");
mztabm_meta.sample[0] = meta_sample;

mztabm.setMetaData(mztabm_meta);
```

***MzTabMSmallMoleculeSectionRow*** The *MzTabMSmallMoleculeSectionRow* class represents one row of the small molecule section, which stores one final result to be reported in terms of a molecule that has been quantified (`https://github.com/HUPO-PSI/mzTab/blob/master/specification_document-releases/2_0-Metabolomics-Release/mzTab_format_specification_2_0-M_release.adoc#63-small-molecule-section`) (Fig. 4.6).

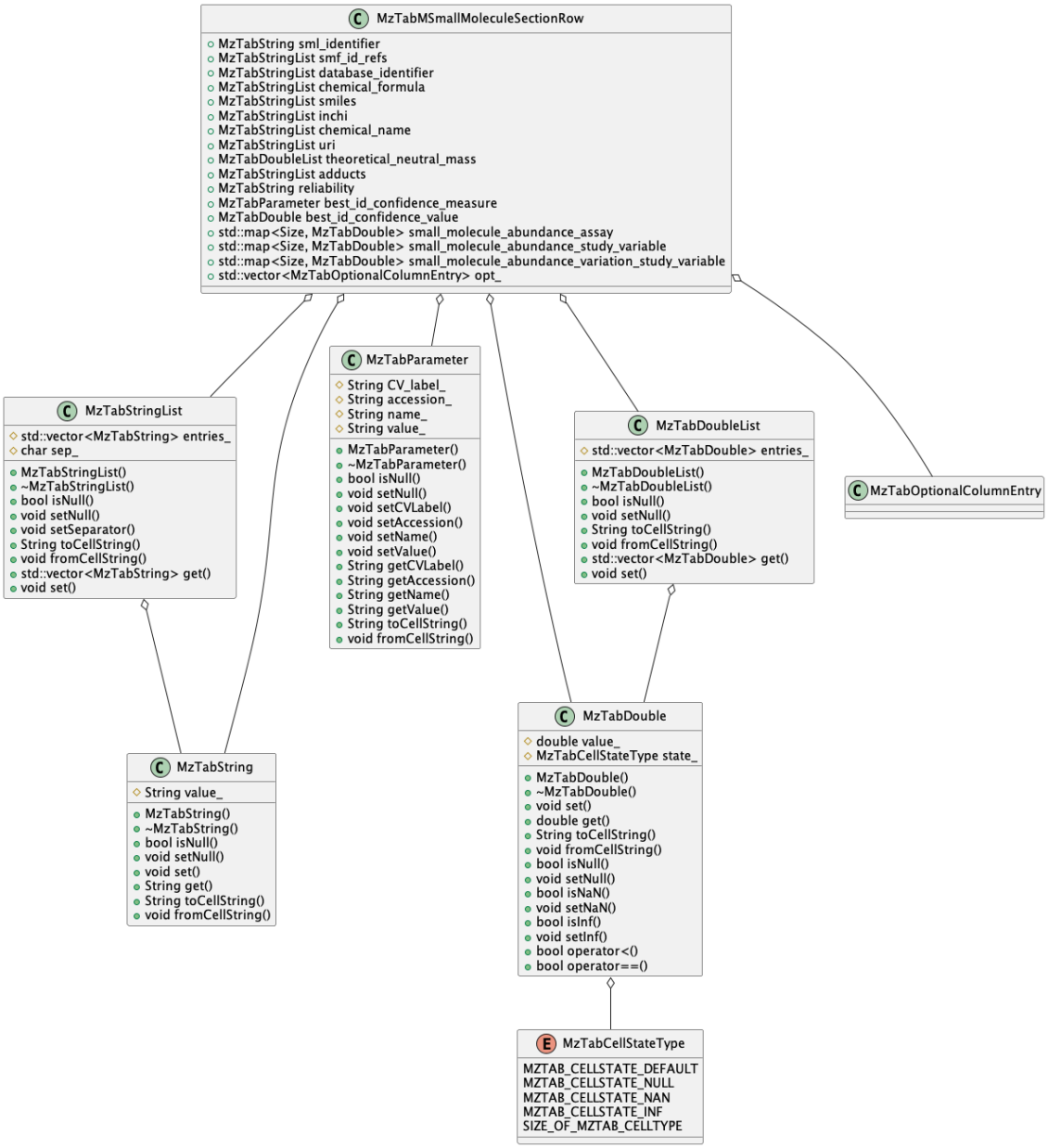


**Figure 4.6: Class Diagram of *SmallMoleculeSectionRow*** The member variables of *MzTabMSmallMoleculeSectionRow* make use of the basic MzTab datatypes (*MzTabString*, *MzTabDouble*, *MzTabParameter*) and their derivatives to store the designated information based on the standard.

A small example is provided to show how data can be added to the *MzTabMSmallMoleculeSectionRow* and the *MzTabM* instance (Listing 4.3).

**Listing 4.3:** Code Snippet Presenting the Usage of *MzTabMSmallMoleculeSectionRow*

```
MzTabM mztabm;

// SML small molecule section row
MzTabMSmallMoleculeSectionRows sml_rows;
MzTabMSmallMoleculeSectionRow sml_row;
sml_row.sml_identifier.fromCellString(1);
sml_row.smf_id_refs.fromCellString("1,2");
sml_row.database_identifier.fromCellString("[HMDB:HMDB0001847]");
sml_row.chemical_formula.fromCellString("[C17H20N4O2]");
sml_row.smiles.fromCellString("[C1=CC=C(C=C1)CCNC(=O)CCNNC(=O)C2=CC=NC=C2]");
sml_row.inchi.fromCellString("[InChI=1S/C17H20N4O2/c22-16(19-12-6-14-4-2-1-3-5-14)
    9-13-20-21-17(23)15-7-10-18-11-8-15/h1-5,7-8,10-11,20H,6,9,12-13H2,(H,19,22)(H
    ,21,23)]");
sml_row.chemical_name.fromCellString("[N-(2-phenylethyl)-3-[2-(pyridine-4-carbonyl
    )hydrazinyl]propanamide]");
sml_row.uri.fromCellString("[http://www.hmdb.ca/metabolites/HMDB0001847]");
vector<MzTabDouble> tnm = {MzTabDouble(312.17)};
sml_row.theoretical_neutral_mass.set(tnm);
sml_row.adducts.fromCellString("[[M+H]1+]");
sml_row.reliability.set("3");
sml_row.best_id_confidence_measure.fromCellString("[MS, MS:1000752, TOPP Software
    ,]");
sml_row.best_id_confidence_value.set(0.4);

MzTabOptionalColumnEntry e;
MzTabString s;
e.first = "SIRIUS_TREE_score";
s.fromCellString("-10.59083");
e.second = s;
sml_row.opt_.emplace_back(e);

sml_rows.emplace_back(sml_row);

mztabm.setMSmallMoleculeSectionRows(sml_rows);
```

***MzTabMSmallMoleculeFeatureSectionRow*** The class *MzTabMSmallMoleculeFeatureSectionRow* represents one row of the small molecule feature section, which holds the quantitative information for a specific feature (`https://github.com/HUPO-PSI/mzTab/blob/master/specification_document-releases/2_0-Metabolomics-Release/mzTab_format_specification_2_0-M_release.adoc#64-small-molecule-feature-smf-section`) (Fig. 4.7).

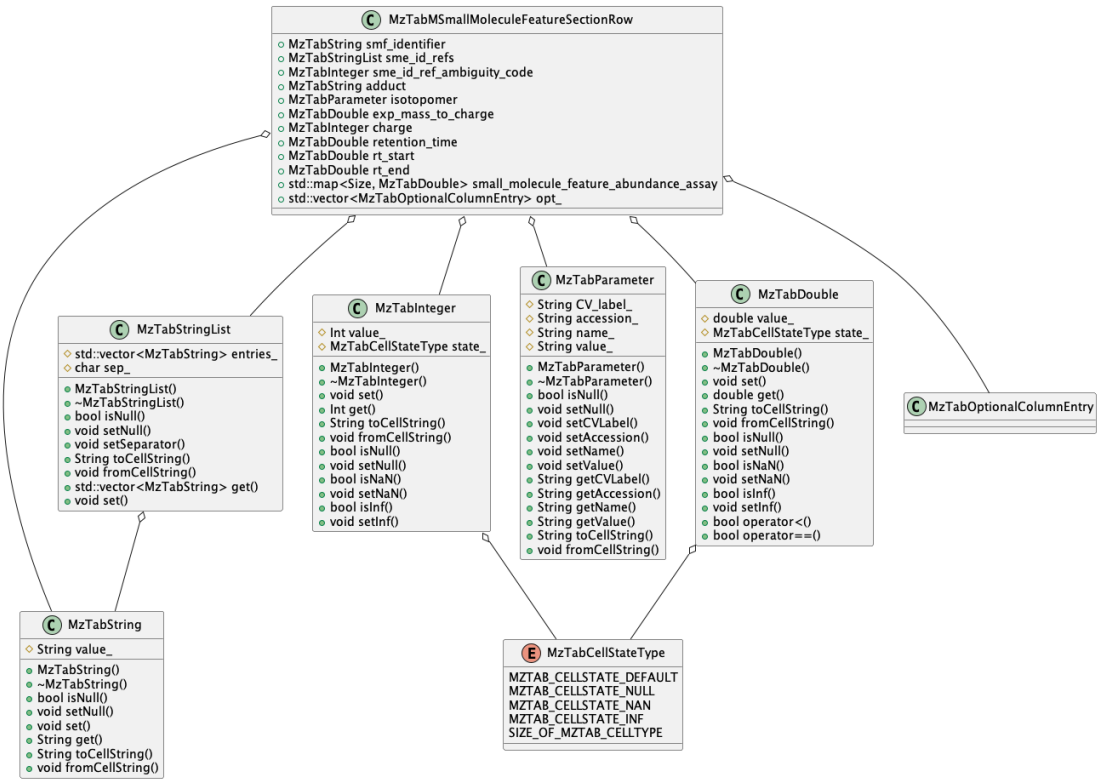


**Figure 4.7: Class Diagram of *MzTabMSmallMoleculeFeatureSectionRow*** The member variables of *MzTabMSmallMoleculeFeatureSectionRow* make use of the basic MzTab datatypes (*MzTabString*, *MzTabInteger*, *MzTabDouble*, *MzTabParameter*) and their derivatives to store the designated information based on the standard.

A small example is provided to show how data can be added to the *MzTabMSmallMoleculeFeatureSectionRow* and the *MzTabM* instance (Listing 4.4).

**Listing 4.4:** Code Snippet Presenting the Usage of *MzTabMSmallMoleculeFeatureSectionRow*

```
MzTabM mztabm;

// SMF Small molecule feature section
MzTabMSmallMoleculeFeatureSectionRows smf_rows;
MzTabMSmallMoleculeFeatureSectionRow smf_row;
smf_row.smf_identifier.fromCellString(1);
smf_row.sme_id_refs.fromCellString("1");
smf_row.sme_id_ref_ambiguity_code.fromCellString("null");
smf_row.adduct.fromCellString("[M+H]1+");
smf_row.isotopomer.setNull(true);
smf_row.exp_mass_to_charge.set(313.1689);
smf_row.charge.set(1);
smf_row.retention_time.set(156.0); // is always in seconds
smf_row.rt_start.set(152.2);
smf_row.rt_end.set(163.4);
smf_rows.emplace_back(smf_row);

mztabm.setMSmallMoleculeFeatureSectionRows(smf_rows);
```

***MzTabMSmallMoleculeEvidenceSectionRow*** The class *MzTabMSmallMoleculeEvidenceSectionRow* represents one row of the small molecule section, which holds identifications evidence of small molecules/features (Fig. 4.8). Each row represents a single result from a database search or other putative identification methods (`https://github.com/HUPO-PSI/mzTab/blob/master/specification_document-releases/2_0-Metabolomics-Release/mzTab_format_specification_2_0-M_release.adoc#65-small-molecule-evidence-sme-section`).

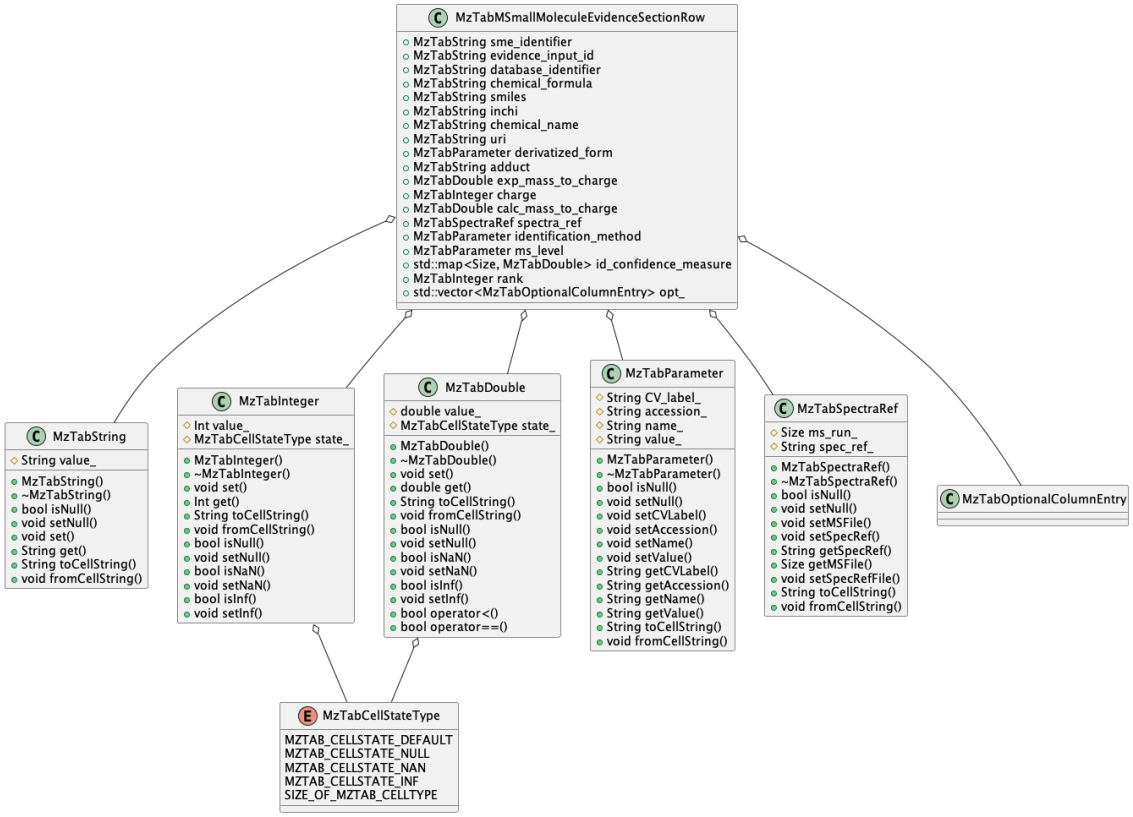


**Figure 4.8: Class Diagram of the *MzTabMSmallMoleculeEvidenceSectionRow*** The member variables of *MzTabMSmallMoleculeEvidenceSectionRow* make use of the basic MzTab datatypes (*MzTabString*, *MzTabInteger*, *MzTabDouble*, *MzTabParameter*, *MzTabSpectraRef*) and their derivatives to store the designated information based on the standard.

A small example is provided to show how data can be added to the *MzTabMSmallMoleculeEvidenceSectionRow* and the *MzTabM* instance (Listing 4.5).

**Listing 4.5:** Code Snippet Presenting the Usage of *MzTabMSmallMoleculeEvidence-SectionRow*

```
MzTabM mztabm;

// SME Small molecule evidence section
MzTabMSmallMoleculeEvidenceSectionRows sme_rows;
MzTabMSmallMoleculeEvidenceSectionRow sme_row;
sme_row.sme_identifier.set(1);
sme_row.evidence_input_id.set("1234.5_156.0");
sme_row.database_identifier.set("HMDB:HMDB0001847");
sme_row.chemical_formula.set("C17H20N4O2");
sme_row.smiles.set("C1=CC=C(C=C1)CCNC(=O)CCNNC(=O)C2=CC=NC=C2");
sme_row.inchi.set("InChI=1S/C17H20N4O2/c22-16(19-12-6-14-4-2-1-3-5-14)
    9-13-20-21-17(23)15-7-10-18-11-8-15/h1-5,7-8,10-11,20H,6,9,12-13H2,(H,19,22)(H
    ,21,23)");
sme_row.chemical_name.set("N-(2-phenylethyl)-3-[2-(pyridine-4-carbonyl)hydrazinyl]
    propanamide");
sme_row.uri.set("http://www.hmdb.ca/metabolites/HMDB0001847");
sme_row.derivatized_form.isNull();
sme_row.adduct.set("[M+H]1+");
sme_row.exp_mass_to_charge.set(313.1689);
sme_row.charge.set(1);
sme_row.calc_mass_to_charge.set(313.1665);
MzTabSpectraRef sp_ref;
sp_ref.setMSFile(1);
sp_ref.setSpecRef("index=5");
sme_row.spectra_ref = sp_ref;
sme_row.identification_method.fromCellString("[MS, MS:1000752, TOPP Software,]");
sme_row.ms_level.fromCellString("[MS, MS:1000511, ms level, 1]");
sme_row.id_confidence_measure[0] = MzTabDouble(123);
sme_row.rank.set(1);

e.first = "SIRIUS_TREE_score";
s.fromCellString("-10.59083");
e.second = s;
sme_row.opt_.emplace_back(e);

sme_rows.emplace_back(sme_row)

mztabm.setMSmallMoleculeEvidenceSectionRows(sme_rows);
```

### MzTab-M serialization

***MzTabMFile*** The class *MzTabMFile* is used for the serialization of the *MzTab-M* data model via the *MzTabM* class to the *MzTab-M* data format (Fig. 4.9). The data model is independent of the actual data format, which keeps the flexibility to implement other non-text-based serialization classes in the future (e.g., to use a repository model). The *MzTabMFile* class has one public store function that allows the serialization to the text-based format.

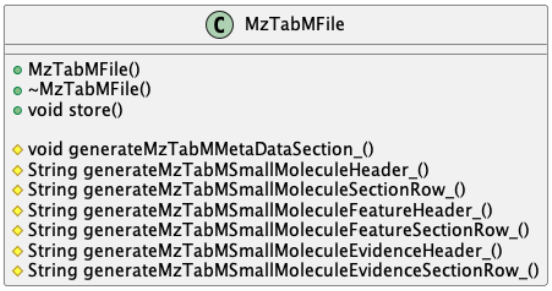


**Figure 4.9: Class Diagram of *MzTabMFile*** The *MzTabMFile* class one public store function that allows the serialization to the text-based format. The protected functions are used to generate the representation of the individual sections and their headers based on the data model.

A small example is provided to show that the information held in the data model (*MzTabM*) instance can be stored in an actual file on the file system (Listing 4.6).

**Listing 4.6:** Code Snippet Presenting the Usage of *MzTabMFile*

```
1
2       FeatureMap feature_map;
3       MzTabM mztabm;
4
5       OMSFile().load(OPENMS_GET_TEST_DATA_PATH("MzTabMFile_input_1.oms"),
            feature_map);
6
7       mztabm = MzTabM::exportFeatureMapToMzTabM(feature_map);
8
9       String mztabm_tmpfile;
10      MzTabMFile().store(mztabm_tmpfile, mztabm);
```

## 4.3 Results

The *MzTab-M* standard was implemented on top of a newly introduced identification (ID) data structure for OpenMS. The ID data structure was conceived and implemented under the lead of Hendrik Weisser, one of the core developers. The first step for the implementation of *MzTab-M* in OpenMS was to incorporate of the ID data structure into *AccurateMassSearch*, a tool used for the identification of analytes based on the MS1 precursor m/z by querying an accurate mass database (i.e., HMBD[32]). This allowed to streamline the development of the *MzTab-M* implementation in OpenMS in correspondence with the ID data structure.
The second step included the implementation of the *MzTab-M* data structure as represented in the specification document. This was realized by adding a *MzTabM* class to the OpenMS library, with the ability to construct a *MzTabM* object by using a *FeatureMap* with the ID information attached. The *FeatureMap* is an OpenMS internal data structure which allows the storage of quantitative feature information, and with the ID data structure is able to also

house ID information.
Lastly a *MztabMFile* class was designed for the serialization to a text file. It uses the provided *MzTab-M* data structure to allow the storage of *MzTab-M* files to disk. Please see the method section for further details in regard to the *MzTabM*, *MzTabMFile,* implementation and usage (Section 4.2). *AccurateMassSearch* is the first and currently only tool supporting *MzTab-M* export in OpenMS.
Another feature which was added with the ID data format was the export of ID extendend *FeatureMap* as .oms file. It is a SQLite based file format that allows the efficient storage of identification and quantification data on disk. We added the export ability to output an *OMSFile* to *AccurateMassSearch*. Conversly, this .oms file can be imported as FeatureMap and we implemented the functionality to convert it directly into an *MzTab-M* object which can be stored as *Mztab-M* file. This basically allows the unconfined export of any metabolomics .oms file into the *MzTab-M* format.
To validate our implementation, we performed a proof-of-concept analysis using our extended *AccurateMassSearch* tool and a sample from the OpenMS metabolomics example data (Section A.2) and were able to successfully export a MzTab-M file, which was further validated using the jmzTab-m validator (Section A.2).

## 4.4 Discussion

We have developed *MzTab-M* for metabolomics data representation and sharing. The standard has been developed in an open process with widespread consultation of different approaches taken in the field and the involvement of software teams from academic research groups and industry. The standard has undergone a rigorous peer review process by both the MSI and PSI to ensure that the resulting standard is of high quality and stability. The standard is expected to remain stable for several years, except for improvements to the documentation and extensions to the CV, allowing research groups and commercial developers to invest time in the implementation. We also encourage other groups interested in standardizing omics data, particularly those using MS (e.g., glycomics), to adopt the *MzTab-M* model/design, CV infrastructure, and associated software. We have implemented and validated the *MzTab-M* format based on the specification into the OpenMS framework to allow for improved interoperability between software used in the Metabolomics community. In terms of *MzTab-M* as export format for metabolomics tools in OpenMS, additional time has to be invested to cover all relevant tools (e.g., *MetaboliteSpectralMatcher*, *SiriusAdapter*, *IdentificationDataConverter*). In addition, to exploit the possibilities of the *MzTab-M* reporting format, a tool with the ability to merge and resolve identification information from various sources (e.g., spectral library search, accurate

mass search, *de-novo* methods), each represented in the evidence section and after scoring the best match would be reported in the summary section could be designed and implemented.

# Chapter 5

# Applied Metabolomics: Food Fingerprinting

This chapter includes partially identical or adapted content with permission from:

---

*Metabolic Fingerprinting: Mass spectrometric determination of the cocoa shell content (Theobroma cacao L.) in cocoa products by HPLC-QTOF-MS*

Nicolas Cain, Oliver Alka, Torben Segelke, Kristian von Wuthenau, Oliver Kohlbacher and Markus Fischer

*Food Chemistry* 298 (2019)

---

A detailed description of the contributions to the project by coauthors is provided in the Appendix C

## 5.1 Introduction

Cacao powder is a crucial ingredient in chocolate products. It is obtained from cacao pods by harvest, fermentation, and processing into cacao powder. Traditionally, the dried and fermented beans are exported from the country of origin to industrialized countries to be roasted and further processed into cocoa products. However, beans are increasingly processed to cocoa mass or other cocoa products in their countries of origin. This geographic shift leads to economic advantages, such as lower transportation weight and costs. However, processing cocoa beans into semi-finished products is a critical quality control problem since standards may not be as stringent as in importing countries. This problem may result from the contamination of the primary product with a higher cacao shell percentage, noticeable by the reduction of

sensory and texture properties not meeting expected quality standards. In addition, stone cells, a specific type of plant cell with highly thickened and lignified cell walls, and fiber structures of the cocoa shell can wear down the production equipment and may lead to undesirable changes in the viscosity and flow properties of the cocoa mass[93,94]. Furthermore, contamination with pollutants, such as mycotoxins, heavy metals, or microorganisms, can occur through the carry-over-process of cocoa shell[95,96,97]. Cocoa shell residues are not entirely avoidable in the course of cocoa bean processing[98]. Here, the roasted beans are broken down, and parts of the shell and the germ are removed by sifting in an airstream or by sieve and vibrating machines. The obtained cocoa extract, which may still contain a small proportion of shell and germ, is the basis for cocoa mass and is further processed into various products. Since controlling the amount of shell in the primary cacao product is not straightforward, the cocoa mass may be stretched by adding shell components for a higher financial gain[99]. If a value of 5% cacao shell and germ is exceeded, stretched goods, inferior beans/raw materials quality, or erroneous processing can be assumed. A robust and efficient method is indispensable to protect the cocoa processing industry from inferior goods and detect food fraud. Up to today, numerous analytical approaches have been developed to quantify the cacao shell content. These include, gravimetric[100,101,102], photometric[103,104,105], liquid chromatographic methods[106] and near-infrared spectroscopy[107]. The current techniques cannot provide robust and reproducible results for samples of different origins, variety, and processing stages. In addition, established methods are based exclusively on determining one or a few compounds[108]. We present a metabolomics-based approach using high-resolution molecular fingerprinting for biomarker identification for developing a new quantification technique for cacao shell content based on different metabolites from various chemical classes.

## 5.2 Materials and Methods

### 5.2.1 Cacao Samples for Biomarker Discovery

Overall, 63 cocoa samples of different processing steps, varieties, origins, and harvest years were obtained. The sample pool consisted of 60 fermented and three unfermented cocoa bean samples. The fermented cocoa bean samples were collected from all commercial relevant countries: Ecuador (10), Ivory Coast (9), Ghana (9), Nigeria (5), Peru (5), Indonesia (4), Cameroon (4), Madagascar (3), Panama (3), Venezuela (2), Sierra Leone (1), Sao Tomé (1), Bolivia (1), Costa Rica (1), Dominican Republic (1), and a mixture of west African countries (1). The unfermented samples originated from Panama. Samples were harvested between 2011 and 2017. For further information about reagents and chemicals, experimental sample preparation and processing, please see Sections A.3.1, A.3.2, A.3.3.

### 5.2.2 Cocoa Calibration Series for the Prediction Model

A cocoa shell calibration series was prepared by mixing defined proportions of cocoa shell and cocoa nibs homogenates. For better miscibility and homogeneity, the cocoa shell homogenate was treated with a ball mill at 3.1 m/s for 5 min. Afterwards, the cocoa shell was mixed with the cocoa nibs in various proportions and homogenized in a mortar with a pestle. A total of eleven different samples were prepared in a concentration range of 0% to 10% of the cocoa shell. For each concentration, approximately 5 g of the mixture was obtained.

### 5.2.3 Data Processing

Two separate data processing strategies were applied to identify and evaluate key metabolites. In the end, the results of both methods were used for comparison and verification of the individual results. After biomarker identification, a regression model was constructed using an in-house R script on a calibration curve with 0% to 10% shell. For the first method (Method 1), the calibration and peak picking of the data sets were carried out using the molecular feature algorithm with DataAnalysis 4.1 (Bruker Daltonics). Next, normalization, time alignment, and peak grouping were performed using ProfileAnalysis 2.1 (Bruker Daltronics). Cocoa nibs and cocoa shells were analyzed separately but evaluated together. Therefore, the peak grouping filter for the minimum signal presence was set to 40% for all samples and the respective groups (nibs and shell) to allow the detection of highly overrepresented features in one group. After level scaling, the resulting bucket list was used for further multivariate analysis, such as Principal Component Analysis (PCA). For the second method (Method 2), the data was processed in three steps (Fig. 5.1). The first step was data management. Here, samples were registered, stored, and initially converted from the mass spectrometry vendor format to the standardized .mzML format using the qportal. In the second step, a computational mass spectrometry analysis was performed using OpenMS 2.3.0[7] and KNIME 3.5.0[51]. In the last step, data analysis entailed filtering, normalization, classification, and feature selection of potential candidates based on the results from the computational MS analysis.

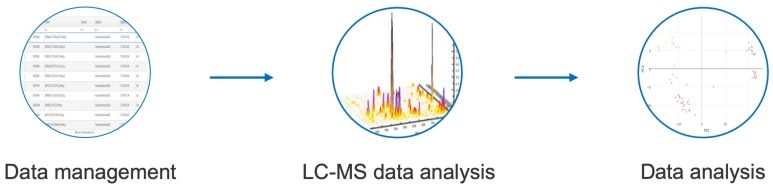


**Figure 5.1: Overview of the Data Processing** The data was processed in three steps. The first step was data management. Here, samples were registered, stored, and initially converted from the mass spectrometry vendor format to the standardized .mzML format using the qportal. In the second step, a computational mass spectrometry analysis was performed using OpenMS and KNIME. In the last step, data analysis entailed filtering, normalization, classification, and feature selection of potential candidates based on the results from the computational MS analysis.

Data management, initial conversion, and centroiding were performed using the qPortal[109] data management system as implemented in the Quantitative Biology Center. Acquired raw data was converted and centroided using msconvert 3.0.9013[61] and the OpenMS PeakPickerHiRes. Computational LC-MS analysis entailed feature detection using the FeatureFinderMetabo (`cacao_feature_extraction_unpolar.toppas`, Fig. 5.2A) and feature linking using the FeatureLinkerUnlabeledQT (`cacao_feature_linking.toppas`, Fig. 5.2B)[62,110]. Feature detection was performed using a noise threshold intensity of 600, a mass error of 5 *ppm*, a maximum mass trace length of 40 seconds, and with report convex hulls enabled. Feature linking was performed with a maximum distance in retention time of 10 *seconds* and a 10 *ppm* error in mass to charge distance. First, the subgroups were linked individually (each group linked over all samples). Then, the resulting individual groups were linked together, while retaining the information of the individual groups. The file name and sample id mapping were constructed using the QBIC barcode (`getNamesConsensusXML_foodomics.py`). Filtering and conversion to a CSV file followed using a KNIME workflow (`ExtractTextConsensusXML_foodomics.knwf`, Fig. 5.2C). The columns rt_cf, mz_cf, and all intensity columns were selected in the column filter node to reduce the CSV file size.

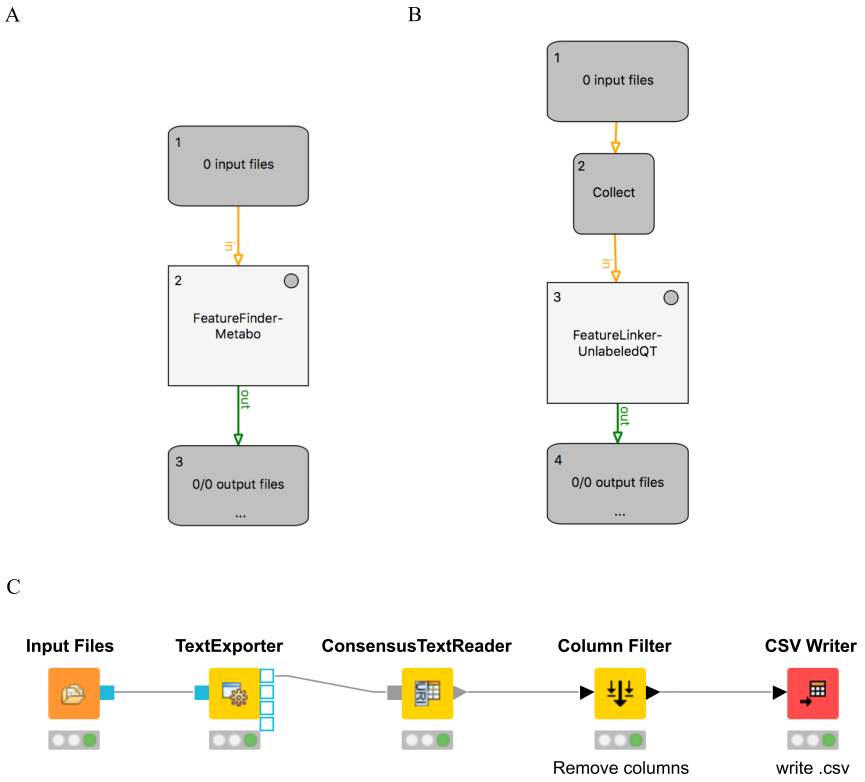


**Figure 5.2: Workflows for the computational mass spectrometry analysis** (A) TOPPAS workflow for feature detection using the FeatureFinderMetabo. Input .mzML files are processed individually and feature detection results are exported as .featureXML file. (B) TOPPAS workflow for feature linking using the FeatureLinkerUnlabeledQT. Input .featureXML/.consensusXML files are collected to a list and feature linking is performed using all files. The results are exported as .consensusXML file. (C) KNIME workflow to export a filtered CSV file from the previously generated .consensusXML file.

Further analysis was performed as depicted in Fig. 5.3 using R 3.4.1 (`cacao_classification_feature_selection.Rmd`).

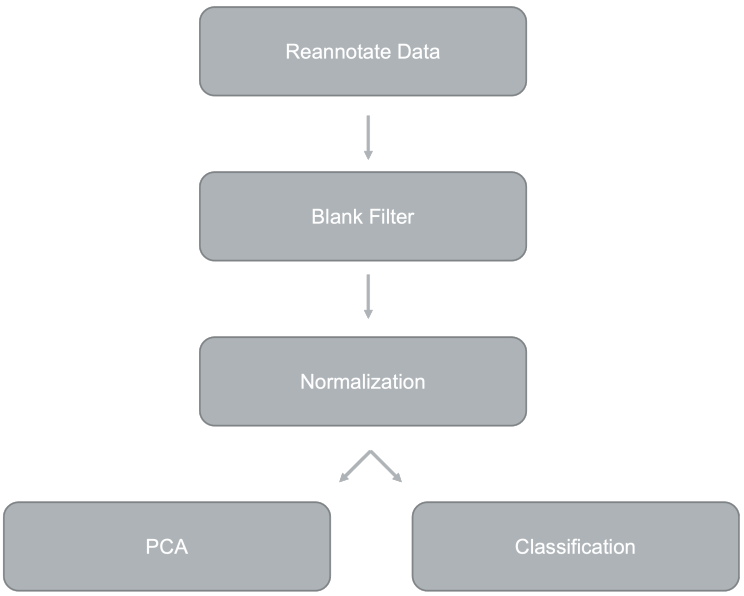


**Figure 5.3: Outline of the Data Analysis** The data analysis of the computational mass spectrometry results was split into multiple parts. First, the data was re-annotated based on a mapping generated from the registered samples based on the data management. Second, a blank filter was performed to reduce the potential feature space. Third, the remaining feature data was normalized. Afterwards, on the one hand, a PCA was performed to get an overview of the separation of the data based on nibs and shell. On the other hand, classification, and feature selection were used to find potential biomarkers.

First, the column names were re-annotated based on the previously generated mapping. Then a blank filter was applied. Features with a 20% higher intensity than the blank and at least 30% of intensity values were retained. The rationale for the intensity difference of 20% in contrast to the blanks was that the features should still be detectable at lower concentrations (e.g., using a reduced sample amount). The second filter criterion states that an intensity value should be present in a minimum of 30% of the samples for a feature, basically filtering for the occurrence of a potential metabolite. Here, features with intensity in either nibs or shell are allowed and expected, but the overall amount of features is reduced depending on the sparsity of the value matrix. As described above, the cacao samples originate from different regions. We assume that differences in the metabolite composition are expected. The aim is to find a method that generalizes as much as possible, indicating that the key metabolites should be detectable in most samples (region independent). After the filtering step, a background normalization was performed. Here, the sample with the most features was used as a reference. The ratio of the other samples relative to the reference was calculated. If the ratio was within a specific outlier range, it was used to normalize the respective sample. A PCA was performed to get an overview of the separation of the data based on nibs and shell samples. In addition, the data set was split into a training (80%) and a test set (20%). The following algorithms were executed using repeated 10-fold cross-validation with five

repetitions. Classification methods, such as random forest (RF[111]), as well as feature selection methods, such as recursive feature elimination (RFE[112]) and all relevant feature selection (Boruta[113]), were applied. The classification method assessed if the data was classifiable, and the feature selection methods were used to find the potential biomarkers. At this point, additional metrics were used to filter for specific marker metabolites. The standard deviation of an analytes intensity based on nibs and shell. The coefficient of variation, which represents the dispersion of data points around the mean (at best, as small as possible). It can be used to assess the intensity differences based on sample origin and, concerning the roast series, the temperature stability. The intensity of an analyte in shell samples had to be higher than 80,000 to ensure that the biomarkers will be detectable in low concentrations samples (based on the experimenters experience regarding this data set). In addition, the amount of missing data over the samples was evaluated, for example, if an analyte was only present in the nibs or shell. After the biomarker and metabolite identification, an additional calibration data set was used to select the prediction model. An inverse sparse partial least squares regression (SPLS) was trained with the data (`cacao_regression.Rmd`). The calibration data set was split into test and training sets based on the replicates. Then the model training for the prediction was performed using fixed parameters previously optimized with ten times repeated 10-fold cross-validation (K=2, eta=0.9, kappa=0.5). Afterwards, the absolute and relative errors were calculated based on the repeated predictions. Workflows and source code used for the analysis are available at (`https://github.com/KohlbacherLab/foodomics-cacao-biomarkeridentification`).

## 5.3 Results

### 5.3.1 Identification, Selection, and Validation of Key Metabolites

Metabolites exclusively contained in the shell are relevant, to detect the concentration of cocoa shells in cocoa products. In the following, the cacao beans after processing are regarded as cacao beans instead of cacao nibs. By performing a selective and comparative analysis of the two groups, bean and shell, a prediction about the cocoa shell content in different cocoa products can be formulated. The identification of key metabolites was performed using two approaches. In the first one, a PCA and a two-sided t-test model using DataAnalysis were used to detect the respective key metabolites. The second method used classification and feature selection for the detection and selection of biomarker candidates.

**Method 1: Dimensionality Reduction, Statistical Validation, and Filtering**

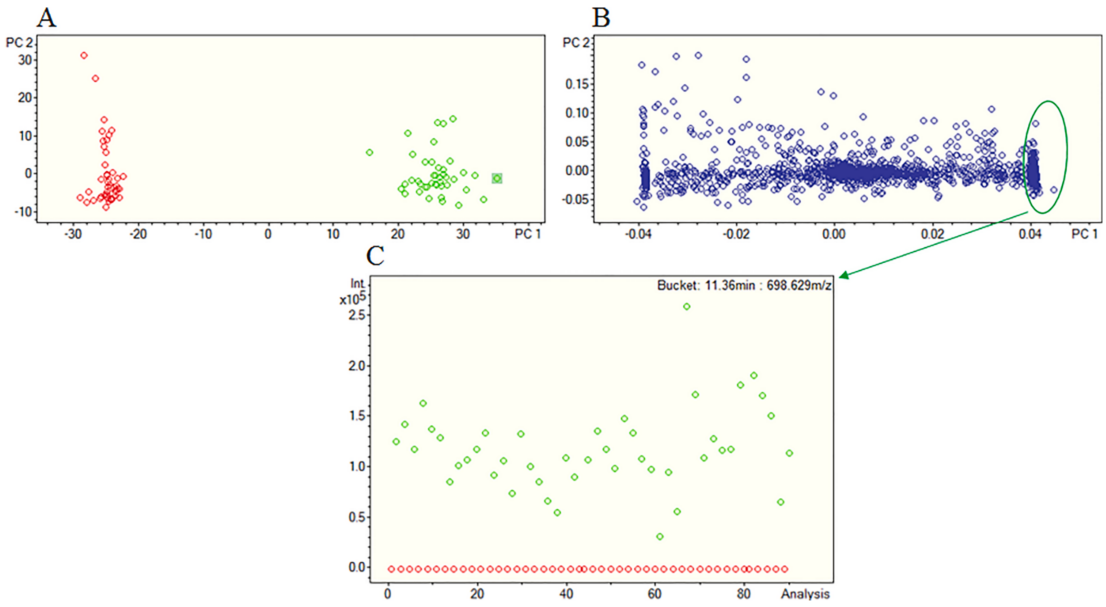


**Figure 5.4: PCA of Bean and Shell Samples** PCA scores (A) and loadings (B) plot of 48 cocoa bean and cocoa shell samples analyzed with the nonpolar positive method displayed PC-1 versus PC-2 (red scores: cocoa bean; green scores: cocoa shell; blue scores: contributing features) and a bucket statistic view (C) of an exemplarily selected potential key metabolite (698.629 m/z and 11.36 min.)

The example in figure 5.4 shows a PCA model based on 48 cocoa beans and the associated cocoa shell samples. In this model, the PC-1 already explains 60% of the variance of the analyzed samples. Since the division between cocoa beans and cocoa shell samples occurs along the PC-1 axis in the scores plot, the potential cocoa shell key metabolites can be found in the region of the highest values (green ellipse) of the PC-1 axis of the loadings plot. For clarification, the figure also shows a bucket statistic view for a feature located in the plot area. The list of potential key metabolites defined by the PCA was supplemented by evaluation using a t-test to ensure significant differences in concentration between metabolites of the two compartments. Only metabolites with an average ratio (cocoa shell/cocoa beans) of $> 5$ and a p-value of $< 2\text{x}10^{-7}$ were considered. A total of 196 potential cocoa shell key metabolites could be identified for the nonpolar positive analysis. Approximately 300 potential cocoa shell key metabolites could be determined for the polar-positive analysis, and about 500 for the polar-negative analysis.
The suitability of the already identified potential key metabolites was further investigated based on the following criteria:

- **Temperature stability criterion:** The concentration of the key metabolites must be independent of the influence of industry-standard roasting temperatures and times.
- **Homogeneity criterion:** The key metabolites must be present in as equal concentrations as possible in each sample, irrespective of the origin, year of harvest, and variety.
- **Fermentation criterion:** Fermentation must not affect the concentration of the key metabolites. E.g., metabolites involved in the metabolism of microbial processes are not suitable. Supposing the concentration of metabolites in unfermented and fermented samples is equal, it can be assumed that the same concentration is also present in samples that have been fermented in other ways.
- **Concentration criterion:** The key metabolites should occur in the highest possible concentration. Therefore, the higher the concentration, the more sensitive and robust the cocoa shell determination will be in chocolate products related to the cocoa mass.

The presented criteria were evaluated as follows: A laboratory-scale roast series, comprised of six differently roasted sample batches was produced and analyzed (Section A.3.2). Then, the obtained area values of the previously defined potential key metabolites were compared with those of unroasted samples of the same batch.

**Table 5.1: Roasting Strategies**

| roasting | temperature [°C] | time [min] |
|---|---|---|
| R0 | not roasted | |
| R1 | 165 | 15 |
| R2 | 165 | 25 |
| R3 | 165 | 35 |
| R4 | 130 | 15 |
| R5 | 130 | 30 |
| R6 | 130 | 45 |
| R7 | 200 | 40 |

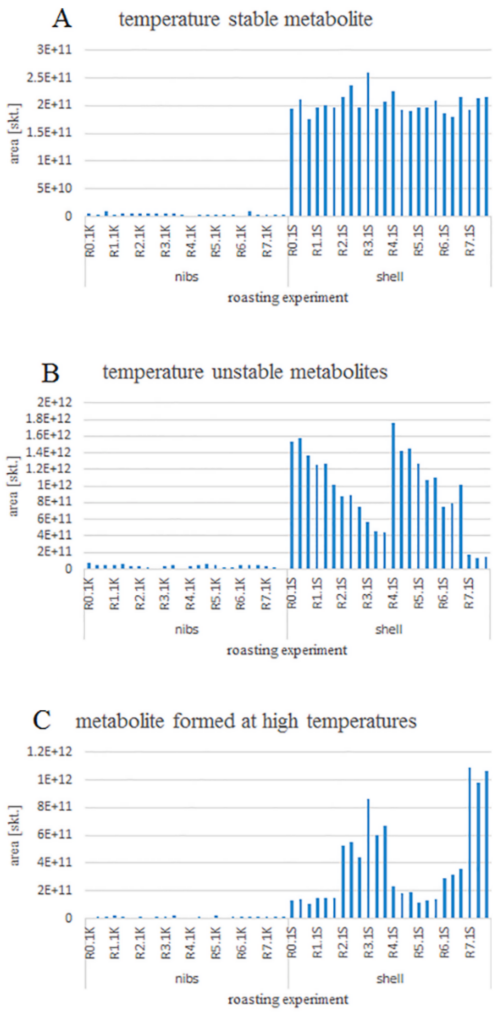


**Figure 5.5: Temperature Stability Criterion (Roest Series)** (A) Example of a temperature stable metabolite ($m/z$ 714.678; RT 13.41 min). (B) Example of a temperature unstable metabolite ($m/z$ 497.416; RT 7.05 min). (C) Example of metabolite formed at high temperatures ($m/z$ 845.824; RT 14.08 min).

Figure 5.5 shows the peak areas of three different potential key metabolites using various roasting strategies (Table 5.1). The compound in Fig. 5.5A is temperature stable. Its peak area does not differ substantially between the individual roasting conditions and the unroasted sample. The metabolite Fig. 5.5B shows a continuous decrease of the peak area to an increase in roasting temperature or roasting duration. Therefore, the compound is temperature unstable. The third compound Fig. 5.5C shows a continuous increase of the peak area to an increase in roasting temperature or roasting duration. This compound arises in the course of the roasting process. Thus, the compounds in diagrams B and C are not suitable for quantifying the cocoa shell content because they depend on the roasting process. Since the different cocoa product

manufacturers often use different roasting processes with different roasting temperatures and times, such behavior is not acceptable. The homogeneity criterion is scrutinized using the coefficient of variation (CV) of the integrated signals of the samples: For the evaluation of the signal variances, the CV as a relative dispersion measure is better for comparing the variance of the metabolites than the standard deviation. Furthermore, the peak areas of the individual compounds were plotted in a bar chart to provide an overview of the content of the individual compounds depending on their origin, variety, and harvest year. Based on this criterion, the content of the respective vital metabolites was analyzed considering 11 different countries of origin, seven different harvest years, as well as bulk and fine flavor cocoa varieties. The homogeneity criterion is satisfied if the CV of one compound is less than 25%, partially met if it is between 25% and 30%, and not met if it is more than 30%. Figure 5.6 shows an example of a potential cocoa shell key metabolite fulfilling the homogeneity criterion.


**Figure 5.6: Homogeneity Criterion** Content of a potential cocoa shell key metabolite in samples of different origin, harvest year, fermentation state, and variety ($m/z$ 686.648; RT 13.13 min. CV=19%) (orange: fermented cocoa beans, yellow: unfermented cocoa beans, green: fermented cocoa shell, red: unfermented cocoa shell;)

The fermentation criterion is verified by comparing the integrated signal areas of fermented with non-fermented cocoa shell samples. If the arithmetic averages of the two sample populations do not differ, a fermentative influence can be excluded. A value of 100 % indicates a non-existent influence of fermentation on the metabolite concentration in question. Since only

three non-fermented samples of one geographical origin could be acquired and analyzed, the amount of data for this criterion is small. Therefore, the values of these samples should be considered indicative. For these reasons, the limits for the fermentation criterion were set less strictly: the tolerated deviation between the content in fermented and non-fermented samples was set to ± 25%. Therefore, all metabolites in the range of 75-125% meet the fermentation criterion. The arithmetic average of the example in figure 5.6 is fulfilled with a score of 93%. For the evaluation of the concentration criterion, the quotient of the signal-to-noise ratios of the cocoa shell and cocoa bean samples were calculated. The criterion is partly met if the quotient is greater than ten and fulfilled if greater than 20. While selecting the key metabolites, the aim was to select those compounds that fulfil as many evaluation criteria as possible.
In summary, for the first method, the previously identified potential key metabolites were examined for temperature stability, fermentation stability, the constant concentration regardless of the origin, year of harvest or variety, and selectively and sensitivity of the detection. After evaluating the metabolites based on the defined criteria, it can be concluded that the potential key metabolites of the nonpolar positive approach are best suited for cocoa shell detection. In the case of the polar methods, only very few compounds showed sufficient temperature stability, and none of the identified temperature-stable compounds showed satisfactory results for all evaluation criteria.

**Method 2: Classification and Feature Selection**

The second method uses classification and feature selection for the detection and selection of biomarker candidates. As the first step, a PCA for each experimental extraction method was generated. It was used to assess the influence of the sample origin and other known metadata. The separability by PCA indicates that further machine learning and feature selection approaches will have no problem separating the two groups. The plots for samples from the polar positive and polar negative extraction do not show a clear separation between bean and shell (Fig. 5.7, 5.8). This behavior could not be attributed to the available metadata (sample type, country of origin, year of harvest, sample processing). The diagram of the polar positive samples may show a batch effect due to the measurement order. This effect is negligible in the case of the following classification analysis. The nonpolar data indicates that the separation between bean and shell is possible independently of the country of origin (Bean: PC1 $< 0$, PC2 $< 0$; Shell PC1 $> 0$; PC2 $< 0$). The outlier (PC2 $> 0$) originates from the roast-series (Fig. 5.9). The progression indicates that potential marker substances must be validated in contrast to roasting and temperature stability (similar to the first method).

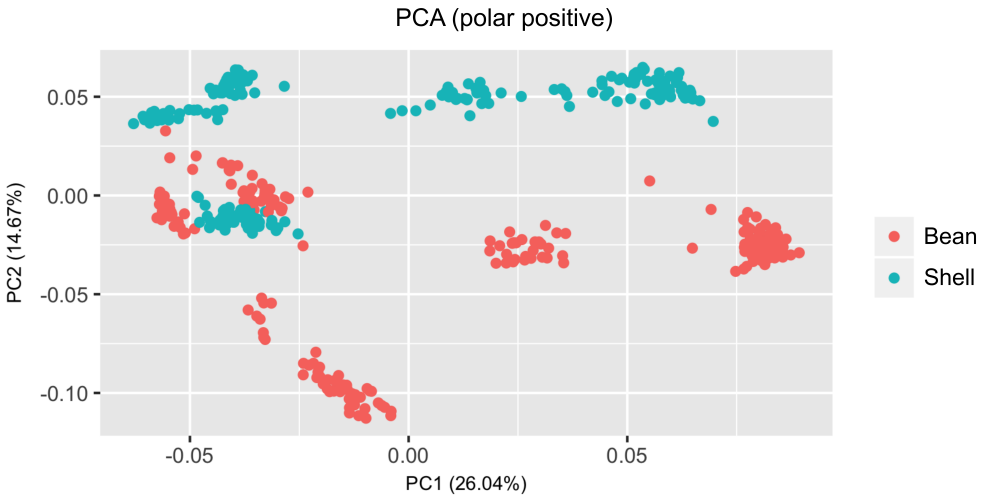


**Figure 5.7: PCA (Polar Positive Extraction Method)** The PCA of the polar positive extraction shows multiple cluster of bean and shell samples which seem to partially separable by PC1 and PC2.

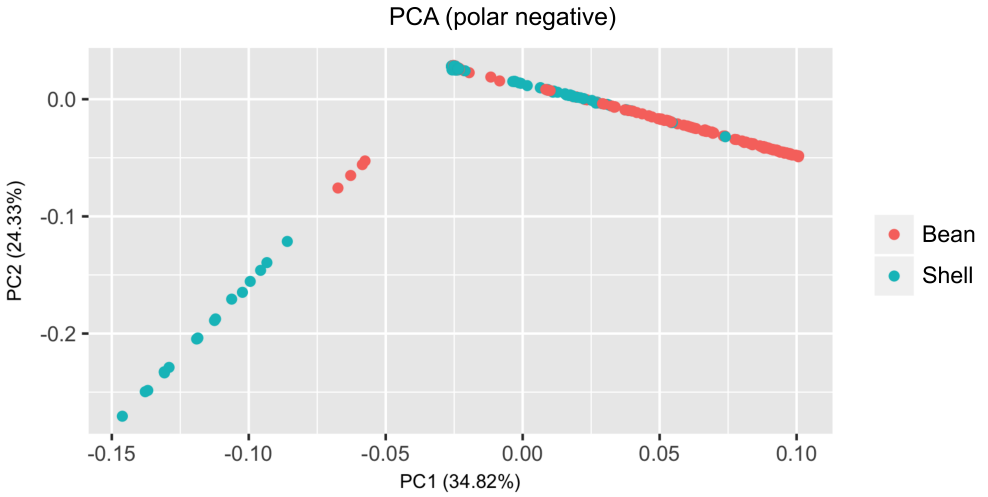


**Figure 5.8: PCA (Polar Negative Extraction Method)** For the polar negative extraction, there is no indication for a clear separation between bean and shell.

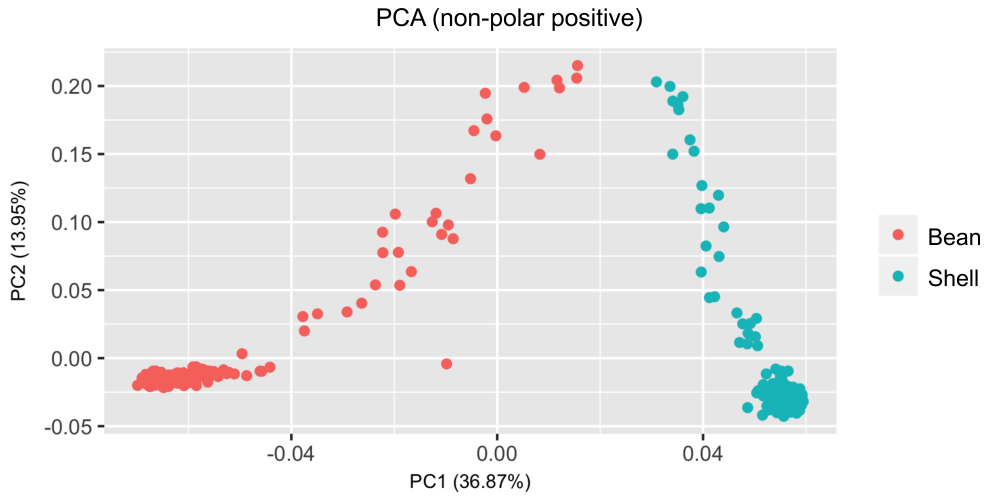


**Figure 5.9: PCA (Nonpolar Positive Extraction Methods)** A clear separability based on PC1 is visible for the nonpolar positive extraction.

As the second step, we used decision trees to assess the separability. The presented examples indicate a similar trend as shown in the PCA examples. For the processed polar samples, positive and negative, multiple features are necessary for a distinct categorization into one of the two groups (Fig. 5.10, 5.11). Such a behavior is not easily translatable into an experimental quantitative method. In contrast, the decision tree of the nonpolar positive method designates that the availability of the substance at $m/z$ 1284.9604 and retention time 694.7 seconds clearly distinguishes the two groups (Fig. 5.12).

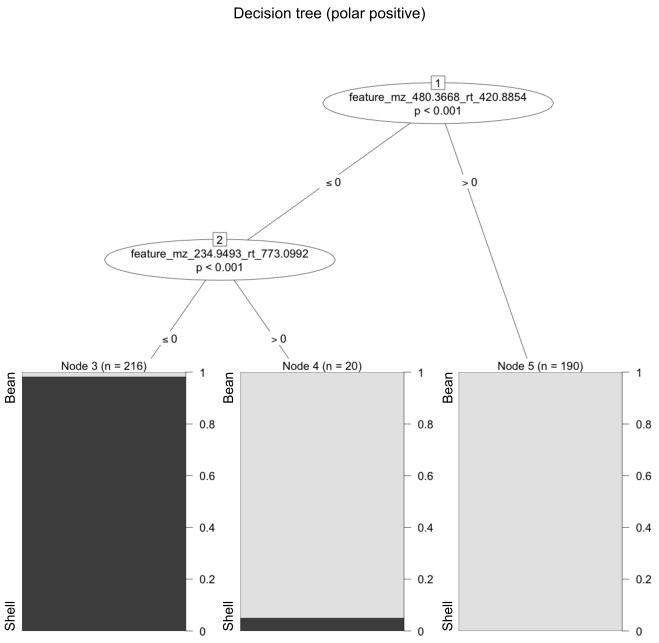


**Figure 5.10: Decision Tree Example for the Polar Positive Extraction Method** For the processed polar positive samples multiple features are necessary for a distinct categorization into one of the two groups.

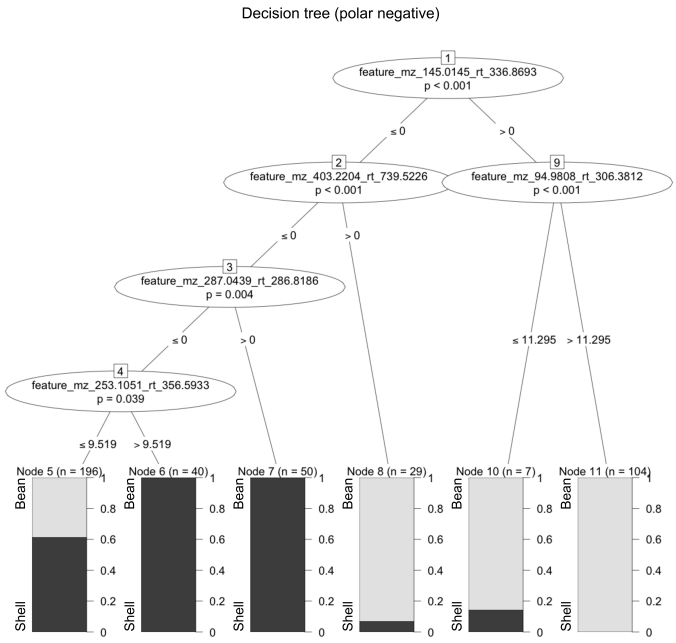


**Figure 5.11: Decision Tree Example for the Polar Negative Extraction Method** For the processed polar negative samples multiple features are necessary for a distinct categorization into one of the two groups.

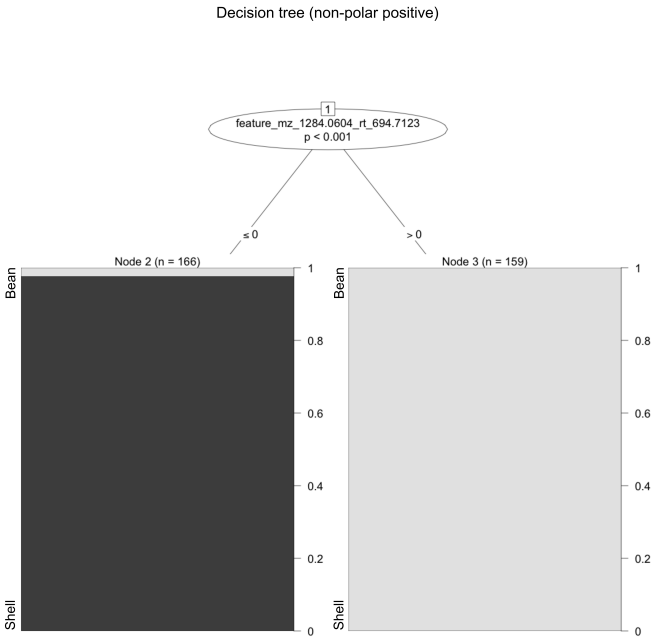


**Figure 5.12: Decision Tree Example for the Non-Polar Positive Extraction Method** The decision tree of the nonpolar positive method designates that the availability of the substance at $m/z$ 1284.9604 and retention time 694.7 seconds clearly distinguishes the two groups.

In consensus with the first analysis method and the findings in the second up to this point, we proceeded with samples from the nonpolar positive extraction method for biomarker selection. For the machine learning algorithms, we aimed for the correct prediction of the two groups, shell and bean, based on the analyte feature space and their intensity. In addition to the classification with RF, we used feature selection algorithms, such as RFE and Boruta.
In short, RF is used to assess which features contribute to the classification but cannot reduce the feature space to the most prominent ones. Here, the feature selection methods come into play. RFE is a method to identify the minimal set of variables by maximizing accuracy. It iteratively removes non-important features until the accuracy of the model drops. In contrast, Boruta tries to assess all relevant features. Boruta works similar to RFE, but it adds randomized copies to the actual feature space, which helps assess a features importance. Boruta stops if no non-important features are left or if the accuracy subsides.

**Table 5.2: Summary of Classification and Feature Selection Results**

| Classification/feature selection | **Training Prediction Accuracy** | **Testing Prediction Accuracy** |
|---|---|---|
| Random Forest | 99.49% | 100.00% |
| Recursive Feature Elimination | 99.88% | 100.00% |
| Boruta with subsequent Random Forest | 99.56% | 100.00% |

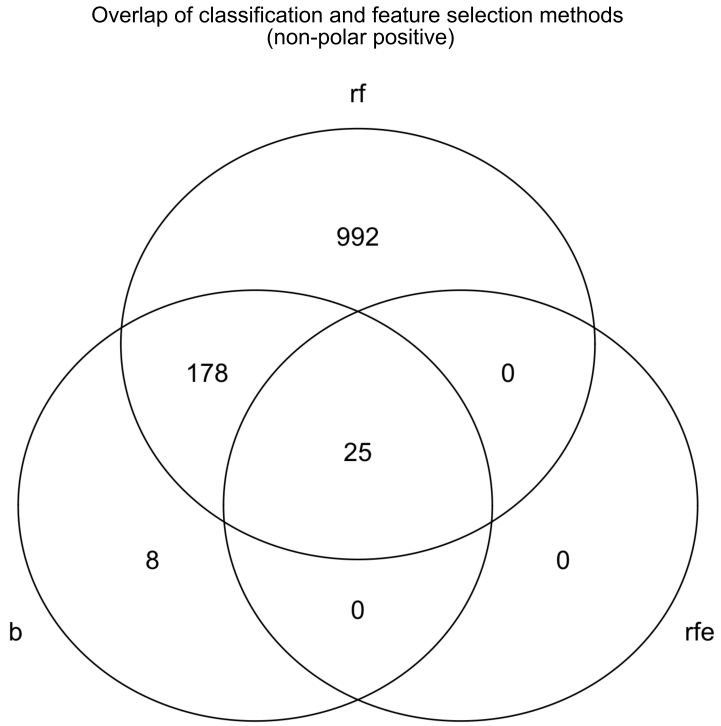


**Figure 5.13: Overlap of Classification and Feature Selection Methods** In total 1203 analytes were used for classification of beans and shell in the nonpolar samples. The set of potential biomarkers were reduced to 211 by the feature selection algorithms. 25 analytes were assessed by RFE in comparison to 211, including the 25 overlapping by Boruta.

The prediction accuracy for all methods was over 99% for the training set (80%) and 100% for the test set (20%) (Table 5.2). The Venn diagram shows the overlap of the classification with the feature selection algorithms.

In total, 1203 analytes were used to classify features to bean and shell in the nonpolar samples. The set of potential biomarkers were reduced to 211 by the feature selection algorithms. Boruta assessed 211 analytes, including the 25 affirmed by RFE (Fig. 5.13). To retain all possible biomarkers, Boruta was trained a second time without splitting the data set. Here, 221 potential biomarkers were detected. Solely focusing on shell biomarkers, 100 analytes remained for

further evaluation. The potential biomarkers were filtered based on criteria similar to those mentioned in the first method. The concentration criterion was fulfilled if the analyte intensity was over 80,000. For exclusivity, the analytes had to have many missing values in the bean, but a low number of missing values in the shell samples, indicating a shell-exclusive metabolite. A low number of missing values ($< 10$) and a low coefficient of variation ($< 40$) in the roast series indicated homogeneity and temperature stability. Applying the filter criteria, 29 possible biomarkers remained.

**Combining Results of Method1 and Method2**

Combining the analysis results from both methods, 10 of the 18 picked metabolites discovered in the first method could be found in the feature selection results and thereby deemed high confidence. In the next step, the actual identification of the biomarker metabolites or their class was performed. First, the key metabolites structural formula was identified by the exact mass, the isotope ratio, and the fragment spectrum. Based on the detected fragments, the potential key metabolites could be assigned to the following substance classes: Fatty acid tryptamines (5), fatty acid serotonin (fatty acid 5-hydroxy-tryptamides) (5), Ceramide derivatives (2), Tocopherol derivatives (3), and triacylglycerols (2). Finally, the validation was performed using standard compounds or class representatives (Section A.3.4).

### 5.3.2 Prediction Model for the Cocoa Shell Content in Cocoa Products

With the relevant and identified biomarkers, the aim was to develop a proof-of-concept method to assess the concentration of shell fragments in a mixture of beans and shell (i.e., representing a primary cacao product). To this end, a cocoa shell calibration series was prepared, analyzed, and evaluated to create a prediction model. The calibration series consisted of a mix of cacao and shell in a concentration of 0% to 10% shell. Five replicates of this series were analyzed. For the generation of a prediction model, the dependency of the variables intensity and concentration were evaluated using linear regression (Fig. 5.14). All metabolites, except of TG(18:2/22:0/18:2) and Cer(d25:0(OH)/18:0(3OH)), follow a linear dependency ($R^2 = 0.88 – 0.95$) and were used to train the prediction model. TG(18:2/22:0/18:2) seems to follow a linear trend but has a high variance over the replicates ($R^2 = 0.64$). In the case of Cer(d25:0(OH)/18:0(3OH)), no linear trend is detected ($R^2 = 0.19$).
An inverse sparse partial least square regression (SPLS) was used for the prediction[114,115], which, after training, allows to directly predict the concentration based on a list of metabolite intensities provided. The model was trained using the centered and scaled calibration data set. It was further validated with five times repeated 10-fold cross-validation. The SPLS model itself reached an $R^2$ of 0.96, root mean square error (RMSE) of 0.78, and a mean absolute error (MAE) of 0.67.

**Table 5.3: Identified Key Metabolites with Selection Criteria** a) Verification of identity against the standard substance; b) Identity verification using a standard substance of the substance class; *) High confidence key metabolites found in both analysis methods

| **metabolite** | **homogeneity criterion CV [%]** | **fermentation criterion [%]** | **concentration criterion** | **$R^2$** |
|---|---|---|---|---|
| α-Tocomonoenolb | 26 | 144 | 40 | 0.991 |
| Heneicosylic acid tryptamide (b),* | 22 | 67 | 13 | 0.989 |
| Docosanoic acid tryptamide (a),* | 8 | 93 | 6 | 0.991 |
| Heneicosylic acid serotonin (b),* | 26 | 68 | 10 | 0.961 |
| Tricosanoic acid tryptamide (b),* | 20 | 79 | 32 | 0.991 |
| Docosanoic acid serotonin (b) | 27 | 124 | 22 | 0.988 |
| Pentacosanoic acid tryptamide (b) | 33 | 84 | 24 | 0.988 |
| Lignoceric acid serotonin (b),* | 44 | 115 | 18 | 0.987 |
| Hexacosanic acid tryptamide (b),* | 33 | 91 | 48 | 0.988 |
| Dihydroceramide (d18:0/16:0) (a),* | 38 | 110 | 24 | 0.988 |
| Pentacosanoic acid serotonin (b) | 24 | 124 | 15 | 0.978 |
| Hexacosanic acid serotonin (b) | 25 | 175 | 6 | 0.989 |
| α-Tocopheryl myristat (b),* | 20 | 101 | 77 | 0.995 |
| α-Tocopheryl palmitate (a) | 19 | 93 | 108 | 0.994 |
| Cer(d25:0(OH)/18:0(3OH)) (b),* | 18 | 232 | 15 | 0.348 |
| TG(18:2/22:0/18:2) (b) | 167 | 34 | 97 | 0.938 |
| TG(22:1/18:0/20:3) (b) | 34 | 118 | 26 | 0,005 |
| Lignoceric acid tryptamide (b) | 21 | 83 | 35 | 0.983 |

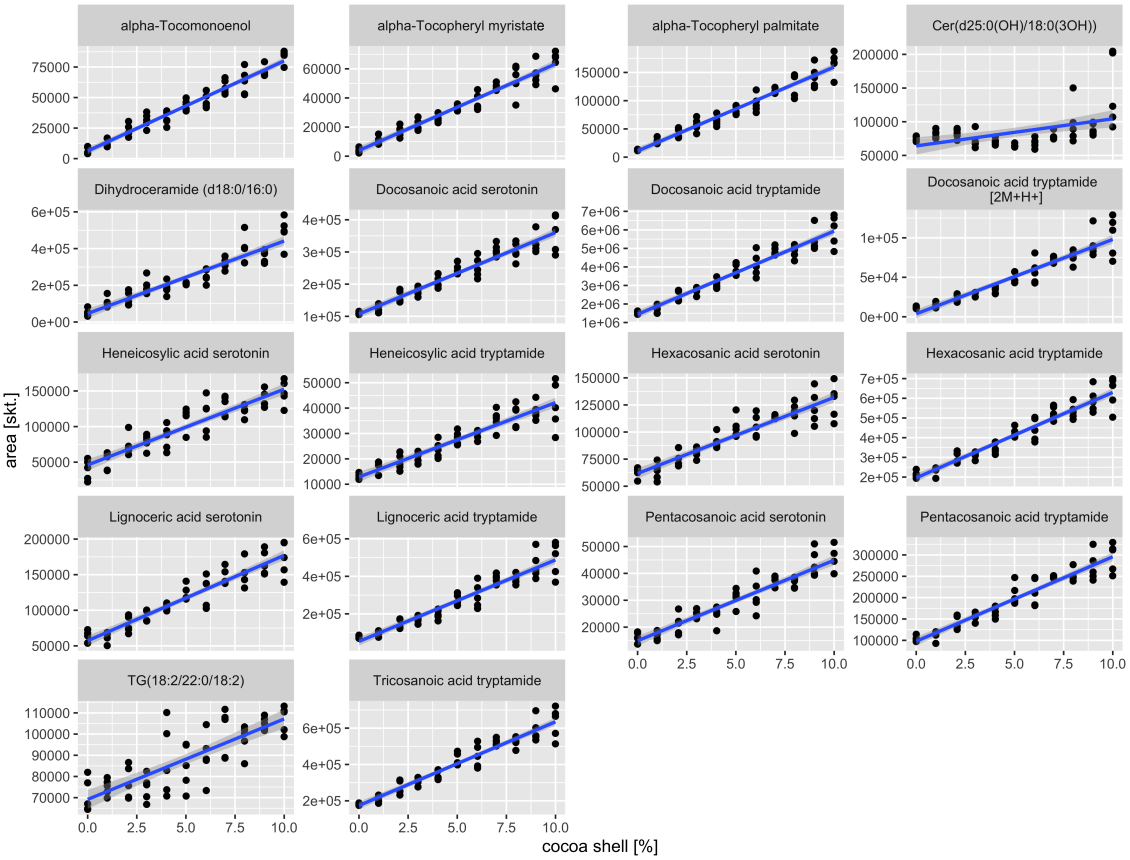


**Figure 5.14: Linear Regression of the Identified Marker Metabolites** All metabolites, except of TG(18:2/22:0/18:2) and Cer(d25:0(OH)/18:0(3OH)), follow a linear dependency ($R^2$ = 0.88 – 0.95) and were used to train the prediction model. TG(18:2/22:0/18:2) seems to follow a linear trend but has a high variance over the replicates ($R^2$ = 0.64). In the case of Cer(d25:0(OH)/18:0(3OH)) no linear trend is recognizable ($R^2$ = 0.19).

Additionally, the data set was partitioned five times iteratively in training (80%) and test set (20%) choosing each replicate. The models were trained on each set using five times repeated ten-fold cross-validation with previously optimized parameters and to ignore possible batch effects. Afterwards, the absolute and relative errors were calculated based on the repeated predictions. The range of the absolute error spans from 0.3% to 1.1% shell. This span correlates with the variance of the replicates observed in Figure 5.14 and the training data set.

**Table 5.4: Validation Using the Test Set (five iterations)**

| target | predicted (mean) | predicted (sd) | absolute error (mean) | absolute error (sd) |
|---|---|---|---|---|
| 0.00 | 0.35 | 0.23 | 0.35 | 0.23 |
| 1.00 | 1.32 | 0.30 | 0.32 | 0.30 |
| 2.08 | 2.37 | 0.69 | 0.60 | 0.34 |
| 3.00 | 3.36 | 0.62 | 0.61 | 0.29 |
| 4.02 | 3.57 | 0.61 | 0.65 | 0.31 |
| 5.01 | 4.99 | 0.43 | 0.34 | 0.19 |
| 6.04 | 5.54 | 1.21 | 1.10 | 0.53 |
| 6.99 | 7.14 | 0.44 | 0.40 | 0.15 |
| 7.98 | 7.77 | 1.15 | 0.98 | 0.44 |
| 9.00 | 8.48 | 1.01 | 0.99 | 0.37 |
| 10.00 | 10.14 | 1.43 | 1.05 | 0.83 |

## 5.4 Discussion

The objective of this chapter was to develop a method for the quality assessment and quantification of cacao shell contamination in an immediate cacao product. Biomarker identification was performed using two methods. The first one used a PCA and t-test to check for distinct key metabolites. Machine learning, classification, and feature selection was the second method. The possible key metabolites were filtered using multiple criteria, such as temperature stability, homogeneity, fermentation, and concentration. The PCA showed that the machine learning algorithms have no issue with the classification and feature selection of the nonpolar samples since the groups were already separable by this method (Fig. 5.4, 5.7, 5.8, 5.9). A similar trend was shown in the decision trees (Fig. 5.10, 5.11, 5.12). Due to this inclination, the biomarker search was narrowed to the nonpolar samples. Comparing the classification, and feature selection algorithms, RF, in contrast to RFE and Boruta, does not restrain the space of potential biomarkers (Fig. 5.13). In addition, the overlap of the machine learning algorithms indicates that *all relevant feature selection* is the method of choice for this analysis since all potential biomarkers are retained. The identified and validated biomarkers showed a linear trend (Fig. 5.14). Therefore, all but two metabolites which showed no sufficient linearity were used as a proof-of-concept to measure the quality (concentration of the shell) of a mixture of bean and shell based on a calibration series using an SPLS method. In summary, the presented study demonstrates the first multiparametric method for cocoa shell detection. The methods used up to now were not able to provide robust and reproducible results for samples of different origins, variety, and different processing stages. Accurate detection of the cocoa shell independent of divergent metadata is possibly based on the identified key metabolites. The prediction of the shell content can be executed by the identified key metabolites using an SPLS based on a calibration data set.

# Chapter 6

# Conclusion and Outlook

The presented thesis covers newly developed methods at different steps of the metabolomics analysis workflow. Continuous advancement of experimental and computational MS methods help to drive the field forward and to reach the goal of a comprehensive identification, quantification, elucidation of function, and interaction of small metabolites in a biological system. At the computational MS step, we encountered the issue of missing FDR estimation and control in metabolomics (Chapter 3). This seriously limits the confidence in reported identifications and quantifications, where manual assessment is still common practice. Therefore, we introduced DIAMetAlyzer, an automated, FDR controlled targeted analysis workflow, enabling a robust FDR estimation for the first time. Our adaptive machine learning approach integrates the MS1 and MS2 spectra signal to optimally separate true signal from noise and provide a well-calibrated FDR estimate. The workflow provides a targeted setting for quantifying known compounds and an untargeted setting to quantify unknown compounds using their m/z patterns. We were able to detect almost twice as many compounds in a targeted setting compared to untargeted deconvolution as is performed by MS-DIAL. In comparison to MetaboDIA, a tool for consensus spectral library building for metabolomics data from DDA experiments, we could quantify 110 additional features. Combining the libraries from MetaboDIA and DIAMetAlyzer increased the number of quantified features by an additional 132. Extracting both known, with prior identification, and unknown compounds, without prior identification, we further improved the number of quantified features by 32%. This can lead to new biological findings, as shown in the analysis of the AMD data set. Using an experimentally specific DDA library based on reference substances allows for accurate identification of compounds and markers from DIA data in low concentrations, facilitating biomarker quantification. The workflow could be extended in the future to work solely on DIA data by using deconvolution methods similar to MS-DIAL and SWATHtoMRM. With the help of fragment annotation and filtering, it would generate high-quality assay libraries from the generated pseudo spectra. It potentially improves the usability, primarily

in *metabolic fingerprinting*, by reducing the cost and experimental resources by removing the DDA measurements. Further, additional identification levels could be incorporated for improvements in identification performance. Here, the fragment-annotated spectra could be exported as a spectral library, and spectral matching could be performed with existing libraries (i.e., NIST, GNPS). All in all, this workflow contributes to the uptake of FDR methods in the metabolomics community.

At the metabolomics analysis workflow reporting step, we looked at a critical aspect of scientific research: representing and sharing analysis results based on the *FAIR* principles (Chapter 4). The *FAIR* principles stand for findable, accessible, interoperable, and reusable[3,4]. In 2014, the human-readable file format *MzTab* was introduced in the proteomics and metabolomics field to allow the distribution of analysis results in a standardized open format[5]. However, in recent years, the limitations of the format regarding metabolomics data have become apparent[6]. Therefore, the Metabolomics Standard Initiative decided to develop an improved interoperable and reusable standard. As a result, we took part in designing the reporting standard *MzTab-M*. The design has undergone a rigorous peer-review process by both the MSI and PSI to ensure that the resulting standard is of high quality and is stable. The standard is expected to remain stable for several years, except for improvements to documentation and extensions to the controlled vocabulary. This allows research groups and commercial developers to invest time in the implementation. We integrated the standard into our OpenMS software framework. Since the current implementation of *Mztab-M* depends on integrating an improved internal identification format provided by Hendrik Weisser, the introduction of the format as output to all metabolomics tools in OpenMS will be an iterative process. In summary, we contributed to the development of the advanced open reporting format *MzTab-M* to improve the interoperability and reusability of results in the metabolomics field.

Lastly, we developed a new method for the post-processing step of a specific application of metabolomics in food chemistry (Chapter 5). It addresses the quality concerns regarding primary cacao products. In recent years, the production of primary cacao products, such as cacao butter, moved from Europe to the cacao-producing countries. This posed challenges to upholding the quality standards of these cacao products in Europe. To this end, a new method needs to be established to allow the quality assessment of such products. We provided the basis for this method using biomarker identification, and machine learning for feature selection to find possible biomarkers. Ten of these were elucidated using two different methods and deemed of high confidence. Most of the putative markers showed a linear trend and could be used via an sparse partial least squares method to determine the shell quantity in a mixture of bean and shell and assess the quality of the product. The presented study demonstrates the first multiparametric method for cocoa shell detection. The methods used up to now were not able to provide robust and reproducible results for samples of different origins, variety, and different processing stages. However, accurate detection of the cocoa shell independent

of divergent metadata is possible based on the identified key metabolites. Furthermore, the prediction of the shell content can be executed by the identified key metabolites using an sparse partial least squares method when based on a calibration data set.

In summary, we developed new methods at different stages of the metabolomics analysis workflow, driving the metabolomics field forward. We are looking forward to where the field is heading and are sure that our contribution in terms of FDR estimation and format specification will improve the comparability, interoperability, and reusability of metabolomics data.

# Bibliography

[1] Kerstin Scheubert, Franziska Hufsky, Daniel Petras, Mingxun Wang, Louis-Félix Nothias, Kai Dührkop, Nuno Bandeira, Pieter C Dorrestein, and Sebastian Böcker. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.*, 8(1):1494, November 2017. v, vii, 2, 19, 25, 28, 29, 51

[2] Xusheng Wang, Drew R Jones, Timothy I Shaw, Ji-hoon Cho, Yuanyuan Wang, Haiyan Tan, Boer Xie, Suiping Zhou, Yuxin Li, and Junmin Peng. Target-Decoy-Based False Discovery Rate Estimation for Large-Scale Metabolite Identification. *J. Proteome Res.*, 17(7):2328–2334, July 2018. v, vii, 2, 19

[3] Mark D Wilkinson, Michel Dumontier, I Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S Grethe, Jaap Heringa, Peter A C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J Lusher, Maryann E Martone, Albert Mons, Abel L Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3:160018, March 2016. v, viii, 2, 54, 102

[4] Mark D Wilkinson, Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S Grethe, Jaap Heringa, Peter A C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J Lusher, Maryann E Martone, Albert Mons, Abel L Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. Addendum: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, 6(1):6, 2019. v, viii, 54, 102

[5] Johannes Griss, Andrew R. Jones, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jürgen Hartler, Gerhard G. Thallinger, Reza M. Salek, Christoph Steinbeck, Nadin Neuhauser, Jürgen Cox, Steffen Neumann, Jun Fan, Florian Reisinger, Qing Wei Xu, Noemi Del Toro, Yasset Pérez-Riverol, Fawaz Ghali, Nuno Bandeira, Ioannis Xenarios, Oliver Kohlbacher, Juan Antonio Vizcaíno, and Henning Hermjakob. The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteomics*, 13(10):2765–2775, 2014. v, viii, 2, 54, 102

[6] Nils Hoffmann, Joel Rein, Timo Sachsenberg, Jürgen Hartler, Kenneth Haug, Gerhard Mayer, Oliver Alka, Saravanan Dayalan, Jake T M Pearce, Philippe Rocca-Serra, Da Qi, Martin Eisenacher, Yasset Perez-Riverol, Juan Antonio Vizcaíno, Reza M Salek, Steffen Neumann, and Andrew R Jones. MzTab-M: A data standard for sharing quantitative results in mass spectrometry metabolomics. *Anal. Chem.*, 91(5):3302–3310, March 2019. v, viii, 2, 54, 102, 161

[7] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E Wolski, Oliver Schilling, Jyoti S Choudhary, Lars Malmström, Ruedi Aebersold, Knut Reinert, and Oliver Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods*, 13(9):741–748, September 2016. 3, 20, 24, 63, 81

[8] S G Oliver, M K Winson, D B Kell, and F Baganz. Systematic functional analysis of the yeast genome. *Trends Biotechnol.*, 16(9):373–8, September 1998. 5

[9] J K Nicholson, J C Lindon, and E Holmes. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica.*, 29(11):1181–9, November 1999. 5

[10] Royston Goodacre, Seetharaman Vaidyanathan, Warwick B Dunn, George G Harrigan, and Douglas B Kell. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.*, 22(5):245–52, May 2004. 5

[11] T Andrew Clayton, John C Lindon, Olivier Cloarec, Henrik Antti, Claude Charuel, Gilles Hanton, Jean-Pierre Provost, Jean-Loïc Le Net, David Baker, Rosalind J Walley, Jeremy R Everett, and Jeremy K Nicholson. Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature*, 440(7087):1073–7, April 2006. 5

[12] Henk den Ouden, Linette Pellis, Guy E H M Rutten, Ilse K Geerars-van Vonderen, Carina M Rubingh, Ben van Ommen, Marjan J van Erk, and Joline W J Beulens. Metabolomic biomarkers for personalised glucose lowering drugs treatment in type 2 diabetes. *Metabolomics*, 12:27, 2016. 5

[13] Daniela Rodrigues, Carmen Jerónimo, Rui Henrique, Luís Belo, Maria de Lourdes Bastos, Paula Guedes de Pinho, and Márcia Carvalho. Biomarkers in bladder cancer: A metabolomic approach using in vitro and ex vivo model systems. *Int. J. cancer*, 139(2):256–68, 2016. 5

[14] Simone Rochfort. Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. *J. Nat. Prod.*, 68(12):1813–20, December 2005. 5

[15] Nicolas Cain, Oliver Alka, Torben Segelke, Kristian von Wuthenau, Oliver Kohlbacher, and Markus Fischer. Food fingerprinting: Mass spectrometric determination of the cocoa shell content (Theobroma cacao L.) in cocoa products by HPLC-QTOF-MS. *Food Chem.*, 298(June):125013, 2019. 5, 162

[16] E Dudley, M Yousef, Y Wang, and W J Griffiths. Targeted metabolomics and mass spectrometry. *Adv. Protein Chem. Struct. Biol.*, 80:45–83, 2010. 5

[17] 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. *Nat. Protoc.*, 2(4):778–791, 2007. 5

[18] Dmitri G. Sitnikov, Cian S. Monnin, and Dajana Vuckovic. Systematic Assessment of Seven Solvent and Solid-Phase Extraction Methods for Metabolomics Analysis of Human Plasma by LC-MS. *Sci. Rep.*, 6(May):1–11, 2016. 6

[19] Leonardo Perez de Souza, Saleh Alseekh, Federico Scossa, and Alisdair R. Fernie. Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research. *Nat. Methods*, 2021. 6

[20] Amanda C. Martin, Alison D. Pawlus, Erin M. Jewett, Donald L. Wyse, Cindy K. Angerhofer, and Adrian D. Hegeman. Evaluating solvent extraction systems using metabolomics approaches. *RSC Adv.*, 4(50):26325–26334, 2014. 6

[21] Dajana Vuckovic. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Anal. Bioanal. Chem.*, 403(6):1523–1548, 2012. 6

[22] Tim J. Causon and Stephan Hann. Review of sample preparation strategies for MS-based metabolomic studies in industrial biotechnology. *Anal. Chim. Acta*, 938:18–32, 2016. 6

[23] Marta Roca, Maria Isabel Alcoriza, Juan Carlos Garcia-Cañaveras, and Agustín Lahoz. Reviewing the metabolome coverage provided by LC-MS: Focus on sample preparation and chromatography-A tutorial. *Anal. Chim. Acta*, 1147:38–55, Feburary 2021. 6

[24] Masamichi Yamashita and John B Fenn. Electrospray ion source. another variation on the free-jet theme. *J. Phys. Chem.*, 88(20):4451–4459, September 1984. 8

[25] Lord Rayleigh. XX. on the equilibrium of liquid conducting masses charged with electricity. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 14(87):184–186, September 1882. 8

[26] Matthias Wilm. Principles of electrospray ionization. *Mol. Cell. Proteomics*, 10(7):1–8, 2011. 8

[27] Ludovic C Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics*, 11(6):O111.016717, June 2012. 12, 24

[28] Bin Zhou, Jun Feng Xiao, Leepika Tuli, and Habtom W Ressom. LC-MS-based metabolomics. *Mol. Biosyst.*, 8(2):470–481, Feburary 2012. 12, 24

[29] J. F. J. Todd. Recommendations for nomenclature and symbolism for mass spectroscopy (including an appendix of terms used in vacuum technology). (Recommendations 1991). *Pure Appl. Chem.*, 63(10):1541–1566, jan 1991. 14

[30] Kjell A. Mortier, Guo-Fang Zhang, Carlos H. Van Peteghem, and Willy E. Lambert. Adduct formation in quantitative bioanalysis: Effect of ionization conditions on paclitaxel. *J. Am. Soc. Mass Spectrom.*, 15(4):585–592, apr 2004. 14

[31] Jian Guo and Tao Huan. Comparison of Full-Scan, Data-Dependent, and Data-Independent acquisition modes in liquid Chromatography-Mass spectrometry based untargeted metabolomics. *Anal. Chem.*, 92(12):8072–8080, June 2020. 17, 24

[32] David S Wishart, Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, Kevin Liang, Rosa Vázquez-Fresno, Tanvir Sajed, Daniel Johnson, Carin Li, Naama Karu, Zinat Sayeeda, Elvis Lo, Nazanin Assempour, Mark Berjanskii, Sandeep Singhal, David Arndt, Yonjie Liang, Hasan Badran, Jason Grant, Arnau Serra-Cayuela, Yifeng Liu, Rupa Mandal, Vanessa Neveu, Allison Pon, Craig Knox, Michael Wilson, Claudine Manach, and Augustin Scalbert. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, 46(D1):D608–D617, January 2018. 17, 38, 44, 76

[33] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, 49(D1):D1388–D1395, January 2021. 17, 146

[34] Manish Sud, Eoin Fahy, Dawn Cotter, Alex Brown, Edward A Dennis, Christopher K Glass, Alfred H Merrill, Jr, Robert C Murphy, Christian R H Raetz, David W Russell, and Shankar Subramaniam. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.*, 35(Database issue):D527–32, January 2007. 17, 38, 44

[35] Stephen E. Stein and Donald R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.*, 5(9):859–866, sep 1994. 18

[36] NIST20: Updates to the NIST tandem and electron ionization spectral libraries. `https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries`. Accessed: 2021-11-4. 18

[37] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crüsemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Cristopher A Boya P, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O'Neill, Enora Briand, Eric J N Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andrés M C Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard O Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutiérrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, and Nuno Bandeira. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.*, 34(8):828–837, August 2016. 18, 38, 128

[38] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, January 2000. 18

[39] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, 34(Database issue):D354–7, January 2006. 18

[40] ChemSpider. `http://www.chemspider.com/`. Accessed: 2021-11-4. 18, 146

[41] Christoph Ruttkies, Emma L Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform.*, 8:3, January 2016. 18, 146

[42] Sebastian Böcker, Matthias C Letzel, Zsuzsanna Lipták, and Anton Pervukhin. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, January 2009. 18

[43] Sebastian Böcker and Florian Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24(16):i49–i55, August 2008. 18

[44] Sebastian Böcker and Kai Dührkop. Fragmentation trees reloaded. *J. Cheminform.*, 8:5, Feburary 2016. 18, 25

[45] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A Aksenov, Alexey V Melnik, Marvin Meusel, Pieter C Dorrestein, Juho Rousu, and Sebastian Böcker. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods*, 16(4):299–302, April 2019. 18, 25, 28, 29

[46] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4(3):207–214, March 2007. 18, 25, 26

[47] Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, 7(1):29–34, January 2008. 19

[48] Lukas Reiter, Oliver Rinner, Paola Picotti, Ruth Hüttenhain, Martin Beck, Mi-Youn Brusniak, Michael O Hengartner, and Ruedi Aebersold. mprophet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods*, 8(5):430–435, May 2011. 19, 26, 29

[49] Johan Teleman, Hannes L Röst, George Rosenberger, Uwe Schmitt, Lars Malmström, Johan Malmström, and Fredrik Levander. DIANA—algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics*, 31(4):555–562, Feburary 2015. 19, 26, 29, 123

[50] George Rosenberger, Isabell Bludau, Uwe Schmitt, Moritz Heusel, Christie L Hunter, Yansheng Liu, Michael J MacCoss, Brendan X MacLean, Alexey I Nesvizhskii, Patrick G A Pedrioli, Lukas Reiter, Hannes L Röst, Stephen Tate, Ying S Ting, Ben C Collins, and Ruedi Aebersold. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods*, 14(9):921–927, September 2017. 19, 26, 29, 123

[51] 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, 2008. 21, 29, 81

[52] Alexander Fillbrunn, Christian Dietz, Julianus Pfeuffer, René Rahn, Gregory A Landrum, and Michael R Berthold. KNIME for reproducible cross-domain analysis of life science data. *J. Biotechnol.*, 261:149–156, November 2017. 21, 29

[53] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A. Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, 46(W1):W537–W544, 2018. 21

[54] Hiroshi Tsugawa, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, and Masanori Arita. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods*, 12(6):523–526, June 2015. 24, 43

[55] Yandong Yin, Ruohong Wang, Yuping Cai, Zhuozhong Wang, and Zheng-Jiang Zhu. DecoMetDIA: Deconvolution of multiplexed MS/MS spectra for metabolite identification in SWATH-MS-Based untargeted metabolomics. *Anal. Chem.*, 91(18):11897–11904, September 2019. 24

[56] Haihong Zha, Yuping Cai, Yandong Yin, Zhuozhong Wang, Kang Li, and Zheng-Jiang Zhu. SWATHtoMRM: Development of High-Coverage targeted metabolomics method using SWATH technology for biomarker discovery. *Anal. Chem.*, 90(6):4062–4070, March 2018. 24

[57] Hannes L Röst, George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M Miladinović, Olga T Schubert, Witold Wolski, Ben C Collins, Johan Malmström, Lars Malmström, and Ruedi Aebersold. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, 32(3):219–223, March 2014. 24, 26, 29

[58] Marcus Ludwig, Markus Fleischauer, Kai Dührkop, Martin A Hoffmann, and Sebastian Böcker. De novo molecular formula annotation and structure elucidation using SIRIUS 4. *Methods Mol. Biol.*, 2104:185–207, 2020. 25

[59] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing, 1995. 26, 46, 47, 128

[60] Olga T Schubert, Ludovic C Gillet, Ben C Collins, Pedro Navarro, George Rosenberger, Witold E Wolski, Henry Lam, Dario Amodei, Parag Mallick, Brendan MacLean, and Ruedi Aebersold. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.*, 10(3):426–441, March 2015. 28, 127

[61] Matthew C Chambers, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jarrett Egertson, Katherine Hoff, Darren Kessner, Natalie Tasman, Nicholas Shulman, Barbara Frewen, Tahmina A Baker, Mi-Youn Brusniak, Christopher Paulse, David Creasy, Lisa Flashner, Kian Kani, Chris Moulding, Sean L Seymour, Lydia M Nuwaysir, Brent Lefebvre, Frank Kuhlmann, Joe Roark, Paape Rainer, Suckau Detlev, Tina Hemenway, Andreas Huhmer, James Langridge, Brian Connolly, Trey Chadick, Krisztina Holly, Josh Eckels, Eric W Deutsch, Robert L Moritz, Jonathan E Katz, David B Agus, Michael MacCoss, David L Tabb, and Parag Mallick. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, 30(10):918–920, October 2012. 28, 38, 82, 127

[62] Erhan Kenar, Holger Franken, Sara Forcisi, Kilian Wörmann, Hans-Ulrich Häring, Rainer Lehmann, Philippe Schmitt-Kopplin, Andreas Zell, and Oliver Kohlbacher. Automated label-free quantification of metabolites from liquid Chromatography–Mass spectrometry data. *Mol. Cell. Proteomics*, 13(1):348–359, January 2014. 28, 82

[63] Chris Bielow, Silke Ruzek, Christian G Huber, and Knut Reinert. Optimal decharging and clustering of charge ladders generated in ESI–MS, 2010. 28

[64] Tobias Bruderer, Emmanuel Varesio, Anita O Hidasi, Eva Duchoslav, Lyle Burton, Ron Bonner, and Gérard Hopfgartner. Metabolomic spectral libraries for data-independent SWATH liquid chromatography mass spectrometry acquisition. *Anal. Bioanal. Chem.*, 410(7):1873–1884, March 2018. 28, 120

[65] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E Wolski, Oliver Schilling, Jyoti S Choudhary, Lars Malmström, Ruedi Aebersold, Knut Reinert, and Oliver Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods*, 13(9):741–748, August 2016. 28, 29

[66] Hannes L Röst, Uwe Schmitt, Ruedi Aebersold, and Lars Malmström. pyOpenMS: A python-based interface to the OpenMS mass-spectrometry algorithm library, 2014. 29

[67] Chih-Chiang Tsou, Dmitry Avtonomov, Brett Larsen, Monika Tucholska, Hyungwon Choi, Anne-Claude Gingras, and Alexey I Nesvizhskii. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods*, 12(3):258–64, 7 p following 264, March 2015. 38, 127

[68] H Paul Benton, Elizabeth J Want, and Timothy M D Ebbels. Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data. *Bioinformatics*, 26(19):2488–2489, October 2010. 38, 127, 128

[69] 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. *Anal. Chem.*, 84(1):283–289, January 2012. 38, 127, 128

[70] Shubham Gupta, Sara Ahadi, Wenyu Zhou, and Hannes Röst. DIAlignR provides precise retention time alignment across distant runs in DIA and targeted proteomics. *Mol. Cell. Proteomics*, 18(4):806–817, April 2019. 38, 128

[71] Zhiqiang Pang, Jasmine Chong, Guangyan Zhou, David Anderson de Lima Morais, Le Chang, Michel Barrette, Carol Gauthier, Pierre-Étienne Jacques, Shuzhao Li, and Jianguo Xia. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res.*, 49(W1):W388–W396, July 2021. 38

[72] Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, and Michael J MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, April 2010. 39

[73] Gengbo Chen, Scott Walmsley, Gemmy C M Cheung, Liyan Chen, Ching-Yu Cheng, Roger W Beuerman, Tien Yin Wong, Lei Zhou, and Hyungwon Choi. Customized consensus spectral library

building for untargeted quantitative metabolomics analysis with data independent acquisition mass spectrometry and MetaboDIA workflow. *Anal. Chem.*, 89(9):4897–4906, May 2017. 43, 44

[74] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47, April 2015. 46, 47

[75] Sabrina L Mitchell, Karan Uppal, Samantha M Williamson, Ken Liu, L Goodwin Burgess, Vilinh Tran, Allison C Umfress, Kelli L Jarrell, Jessica N Cooke Bailey, Anita Agarwal, Margaret Pericak-Vance, Jonathan L Haines, William K Scott, Dean P Jones, and Milam A Brantley, Jr. The carnitine shuttle pathway is altered in patients with neovascular Age-Related macular degeneration. *Invest. Ophthalmol. Vis. Sci.*, 59(12):4978–4985, October 2018. 48

[76] Xiao-Wen Hou, Ying Wang, and Chen-Wei Pan. Metabolomics in Age-Related macular degeneration: A systematic review. *Invest. Ophthalmol. Vis. Sci.*, 61(14):13, December 2020. 48

[77] Andreas Reichenbach and Andreas Bringmann. Purinergic signaling in retinal degeneration and regeneration. *Neuropharmacology*, 104:194–211, May 2016. 48

[78] Wei Zhu, Yi-Fang Meng, Qian Xing, Jian-Jun Tao, Jiong Lu, and Yan Wu. Identification of lncRNAs involved in biological regulation in early age-related macular degeneration. *Int. J. Nanomedicine*, 12:7589–7602, October 2017. 48

[79] Bum-Joo Cho, Jang Won Heo, Tae Wan Kim, Jeeyun Ahn, and Hum Chung. Prevalence and risk factors of age-related macular degeneration in korea: the korea national health and nutrition examination survey 2010-2011. *Invest. Ophthalmol. Vis. Sci.*, 55(2):1101–1108, Feburary 2014. 48

[80] Zuhal Yildirim, Nil Irem Ucgun, and Filiz Yildirim. The role of oxidative stress and antioxidants in the pathogenesis of age-related macular degeneration. *Clinics*, 66(5):743–746, 2011. 49

[81] Bénédicte M J Merle, Pascale Benlian, Nathalie Puche, Ana Bassols, Cécile Delcourt, Eric H Souied, and Nutritional AMD Treatment 2 Study Group. Circulating omega-3 fatty acids and neovascular age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.*, 55(3):2010–2019, March 2014. 49

[82] Nicolas G Bazan. Neuroprotectin d1-mediated anti-inflammatory and survival signaling in stroke, retinal degenerations, and alzheimer's disease. *J. Lipid Res.*, 50 Suppl:S400–5, April 2009. 49

[83] John Paul SanGiovanni and Emily Y Chew. The role of omega-3 long-chain polyunsaturated fatty acids in health and disease of the retina. *Prog. Retin. Eye Res.*, 24(1):87–138, January 2005. 49

[84] Emma L. Schymanski, Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P. Singer, and Juliane Hollender. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.*, 48(4):2097–2098, Feburary 2014. 50, 62

[85] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.*, 28(1):31–36, Feburary 1988. 58, 60, 62

[86] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. InChI - the worldwide chemical structure identifier standard. *J. Cheminform.*, 5(1):7, December 2013. 58, 60, 62

[87] Lloyd W Sumner, Alexander Amberg, Dave Barrett, Michael H Beale, Richard Beger, Clare A Daykin, Teresa W-M Fan, Oliver Fiehn, Royston Goodacre, Julian L Griffin, Thomas Hankemeier, Nigel Hardy, James Harnly, Richard Higashi, Joachim Kopka, Andrew N Lane, John C Lindon, Philip Marriott, Andrew W Nicholls, Michael D Reily, John J Thaden, and Mark R Viant. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, 3(3):211–221, September 2007. 62

[88] Gerhard Mayer, Luisa Montecchi-Palazzi, David Ovelleiro, Andrew R Jones, Pierre-Alain Binz, Eric W Deutsch, Matthew Chambers, Marius Kallhardt, Fredrik Levander, James Shofstahl, Sandra Orchard, Juan Antonio Vizcaíno, Henning Hermjakob, Christian Stephan, Helmut E Meyer, Martin Eisenacher, and HUPO-PSI Group. The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database (Oxford).*, 2013:bat009, 2013. 63

[89] Luisa Montecchi-Palazzi, Samuel Kerrien, Florian Reisinger, Bruno Aranda, Andrew R. Jones, Lennart Martens, and Henning Hermjakob. The PSI semantic validator: A framework to check MIAPE compliance of proteomics data. *Proteomics*, 9(22):5112–5119, November 2009. 63

[90] 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. *Anal. Chem.*, 78(3):779–787, Feburary 2006. 63

[91] Jürgen Hartler, Alexander Triebl, Andreas Ziegl, Martin Trötzmüller, Gerald N Rechberger, Oana A Zeleznik, Kathrin A Zierler, Federico Torta, Amaury Cazenave-Gassiot, Markus R Wenk, Alexander Fauland, Craig E Wheelock, Aaron M Armando, Oswald Quehenberger, Qifeng Zhang, Michael J O Wakelam, Guenter Haemmerle, Friedrich Spener, Harald C Köfeler, and Gerhard G Thallinger. Deciphering lipid structures based on platform-independent decision rules. *Nat. Methods*, 14(12):1171–1174, December 2017. 63

[92] Kenneth Haug, Reza M. Salek, Pablo Conesa, Janna Hastings, Paula de Matos, Mark Rijnbeek, Tejasvi Mahendraker, 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 Res.*, 41(D1):D781–D786, January 2013. 63

[93] Steve T Beckett. *Industrial Chocolate Manufacture and Use*. John Wiley & Sons, September 2011. 80

[94] S T Beckett. *The Science of Chocolate*. Royal Society of Chemistry, 2019. 80

[95] S Amézqueta, E González-Peñas, M Murillo, and A López de Cerain. Occurrence of ochratoxin a in cocoa beans: effect of shelling. *Food Addit. Contam.*, 22(6):590–596, June 2005. 80

[96] M Raters and R Matissek. Study on distribution of mycotoxins in cocoa beans. *Mycotoxin Res.*, 21(3):182–186, September 2005. 80

[97] Marina V Copetti, Beatriz T Iamanaka, Melanie A Nester, Priscilla Efraim, and Marta H Taniwaki. Occurrence of ochratoxin a in cocoa by-products and determination of its reduction during chocolate manufacture. *Food Chem.*, 136(1):100–104, January 2013. 80

[98] Bernard Minifie. *Chocolate, Cocoa and Confectionery: Science and Technology*. Springer Science & Business Media, December 2012. 80

[99] Emmanuel Ohene Afoakwa. *Chocolate Science and Technology*. John Wiley & Sons, Nashville, TN, 2 edition, June 2016. 80

[100] W L Dubois and C I Lott. Determination of cocoa shells in cocoa powder. *J. Ind. Eng. Chem.*, 3(4):251–252, April 1911. 80

[101] Arthur W Knapp and Basil G McLellan. The estimation of cacao shell. *Analyst*, 44(514):2, 1919. 80

[102] Julian L Baker and H F E Hulton. Estimation of shell in cocoa and cacao products. *Analyst*, 43(507):197, 1918. 80

[103] A Fincke and H Sacher. Investigations into testing the purity of cocoa butter and chocolate fats. part 6: Quantitative evaluation of the colour reaction with P-Dimethylaminobenzaldehyde for evidence of cocoa shell fat. *Süsswaren*, 7:428–431, 1963. 80

[104] H Szeląg and W Zwierzykowski. Evaluation of behenic acid tryptamide in cocoa fat on the basis of blue value determinations. *Nahrung*, 32(3):285–290, 1988. 80

[105] Michael Münch and P Schieberle. A sensitive and selective method for the quantitative determination of fatty acid tryptamides as shell indicators in cocoa products. *Zeitschrift für Lebensmitteluntersuchung und -Forschung A*, 208(1):39–46, January 1999. 80

[106] Katrin Janßen and Reinhard Matissek. Fatty acid tryptamides as shell indicators for cocoa products and as quality parameters for cocoa butter. *Eur. Food Res. Technol.*, 214(3):259–264, December 2001. 80

[107] Maribel Alexandra Quelal-Vásconez, María Jesús Lerma-García, Édgar Pérez-Esteve, Alberto Arnau-Bonachera, José Manuel Barat, and Pau Talens. Fast detection of cocoa shell in cocoa powders by near infrared spectroscopy and multivariate analysis. *Food Control*, 99:68–72, May 2019. 80

[108] Katja Dettmer, Pavel A Aronov, and Bruce D Hammock. Mass spectrometry-based metabolomics, 2007. 80

[109] Christopher Mohr, Andreas Friedrich, David Wojnar, Erhan Kenar, Aydin Can Polatkan, Marius Cosmin Codrea, Stefan Czemmel, Oliver Kohlbacher, and Sven Nahnsen. qportal: A platform for data-driven biomedical research. *PLoS One*, 13(1):e0191603, January 2018. 82

[110] Hendrik Weisser, Sven Nahnsen, Jonas Grossmann, Lars Nilse, Andreas Quandt, Hendrik Brauer, Marc Sturm, Erhan Kenar, Oliver Kohlbacher, Ruedi Aebersold, and Lars Malmström. An automated pipeline for high-throughput label-free quantitative proteomics. *J. Proteome Res.*, 12(4):1628–1644, April 2013. 82

[111] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. 85

[112] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1):389–422, January 2002. 85

[113] Miron B. Kursa and Witold R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010. 85

[114] Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B Stat. Methodol.*, 72(1):3–25, January 2010. 96

[115] Alejandro C Olivieri. *Introduction to Multivariate Calibration: A Practical Approach*. Springer, Cham, 2018. 96

[116] Jamie Sherman, Matthew J McKay, Keith Ashman, and Mark P Molloy. How specific is my SRM?: The issue of precursor and product ion redundancy. *Proteomics*, 9(5):1120–1123, March 2009. 119

[117] Jamie Sherman, Matthew J McKay, Keith Ashman, and Mark P Molloy. Unique ion signature mass spectrometry, a deterministic method to assign peptide identity. *Mol. Cell. Proteomics*, 8(9):2051–2062, September 2009. 119

[118] Hannes Röst, Lars Malmström, and Ruedi Aebersold. A computational tool to detect and avoid redundancy in selected reaction monitoring. *Mol. Cell. Proteomics*, 11(8):540–549, August 2012. 119

[119] Premy Shanthamoorthy, Adamo Young, and Hannes Röst. Analyzing assay specificity in metabolomics using unique ion signature simulations. *Anal. Chem.*, 93(33):11415–11423, August 2021. 119

[120] 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, July 2010. 128

[121] Eugene Melamud, Livia Vastag, and Joshua D Rabinowitz. Metabolomic analysis and visualization engine for LC-MS data. *Anal. Chem.*, 82(23):9818–9826, December 2010. 128

[122] Oliver Alka, Premy Shanthamoorthy, Michael Witting, Karin Kleigrewe, Oliver Kohlbacher, and Hannes L. Röst. Diametalyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics. *Nature Communications*, 13:1347, 3 2022. 161

# Appendix A

# Supplementary Information

## A.1 Automated, FDR Controlled Targeted Analysis for Data-Independent Acquisition in Metabolomics

### A.1.1 Dilution Series

Concentration of the 250 pesticides over the 1:4 dilution series (Table A.1). The concentration of the metabolite is dependent on its molecular mass, which ranges from 141 to 874 g/mol. Using an injection volume of 5 µL, the concentration ranges from the highest concentration (step1 - min: 5717.501 fmol/µL, max 35460.666 fmol/µL) to the lowest concentration (step 10 - min: 0.022 fmol/µL, max 0.135 fmol/µL) covering 5 orders of magnitude (Fig. A.1).

**Table A.1: Preparation of the 1:4 Dilution Series of the Pesticide Mix in Blood Plasma Measured via SWATH Acquisitions**

| Step | Dilution | Previous dilution (µL) | Plasma matrix (µL) | Replicates |
|---|---|---|---|---|
| 1 | 1 | - | - | 3 |
| 2 | 4 | 25 | 75 | 3 |
| 3 | 16 | 25 | 75 | 3 |
| 4 | 64 | 25 | 75 | 3 |
| 5 | 256 | 25 | 75 | 3 |
| 6 | 1024 | 25 | 75 | 3 |
| 7 | 4096 | 25 | 75 | 3 |
| 8 | 16384 | 25 | 75 | 3 |
| 9 | 65536 | 25 | 75 | 3 |
| 10 | 262144 | 25 | 75 | 3 |

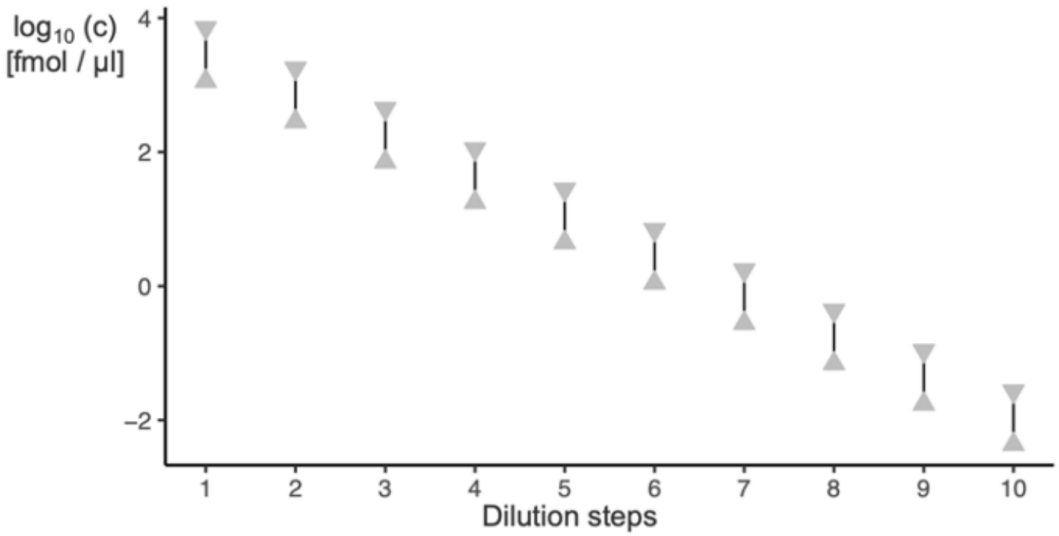


**Figure A.1: Concentration of the Pesticides over the Dilution Series** The concentration gradient over the dilution series covers 5 orders of magnitude and depends on the molecular mass of the pesticide. Covering minimum 5717.501 fmol/µL (step 1) to 0.022 fmol/µL (step 10) and maximum 35460.666 fmol/µL (step 1) to 0.135 fmol/µL (step 10) over the 1:4 dilution series.

### A.1.2 Variable SWATH Windows

The variable SWATH windows were assessed based on the plasma matrix using the SWATH Variable Window Calculator (SCIEX). The target number of windows was set to 8, lower and upper m/z limit to 100 and 900, respectively. *Round bin edges to x figures* were set to 1, and a window overlap of 1 Da was chosen. In addition, the minimum window was set to 3 and CES to 5. With these settings, the windows were determined (Table A.2).

**Table A.2: Variable SWATH Windows Assessed using SWATH Variable Window Calculator Based on the Plasma Matrix**

| Start m/z | End m/z |
|---|---|
| 99.5 | 157.7 |
| 156.7 | 242.6 |
| 241.6 | 370.3 |
| 369.3 | 465.9 |
| 464.9 | 511.7 |
| 510.7 | 557.5 |
| 556.5 | 627.7 |
| 626.7 | 886.1 |

### A.1.3 Optimizing for Unique Identifications

To investigate the effects of assay redundancy and specificity within our assay library we first used computational models to calculate non redundant theoretical assays, also known as unique ion signatures (UIS), for a given metabolomic background[116,117,118]. UISn is defined as a set of top n m/z values (precursor and fragment m/z) that maps exclusively to one metabolite in the metabolome to be analyzed. For this analysis, we simulated classic MS methods (MS1, MRM, SWATH) using the assay library and NIST 17 LC/MS as a combined background with varying mass accuracy for both the precursor m/z window (MS1) and the fragment m/z window (MS2) to compare differences between UIS1, UIS2 and UIS3. This analysis was performed for data acquired in both low (20-50 eV) and high (50-80 eV) ranges of collision energy. This analysis demonstrates that SWATH outperforms MS1-only analyses while performing comparably to MRM-based analyses when using high accuracy fragment ion information only (33.2%, 56.1% and 62.8% unambiguous compounds respectively for MS1-only, SWATH Variable Windows/25 ppm and MRM UIS3 (Fig. A.2a). However, with increased MS1 accuracy, SWATH outperforms MRM-based analyses (with a few transitions) due to its capability of extracting both high-resolution MS1 precursor ion traces as well as high-resolution fragment ion traces for any analyte of interest (93.4% unambiguous compounds for SWATH 25 ppm/25 ppm UIS3, simulating both MS1 and MS2 XIC analysis (Fig. A.2). Based on these overall results and our previous study[119], scoring is based on both MS1 and MS2 information, in addition to retention time and fragment ion relative intensity.

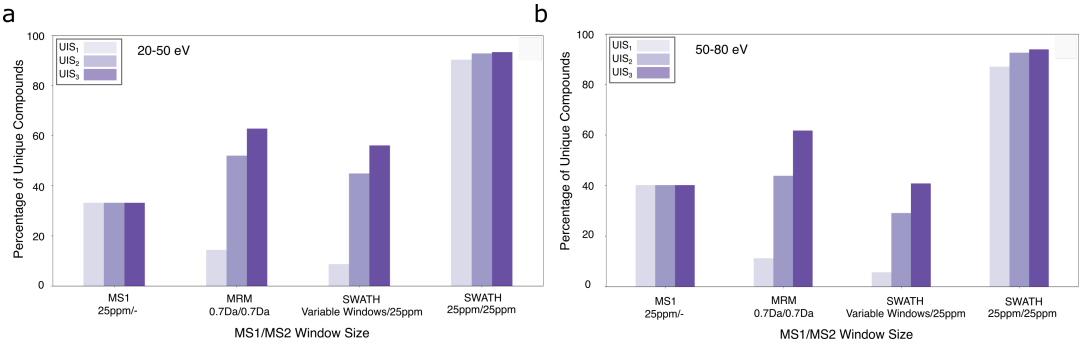


**Figure A.2: Percentage of Unique Compounds for Different MS Methods by Comparison of UIS1, UIS2 and UIS3** Using the assay library and the NIST 17 LC/MS database as background, simulations were conducted for each analyte (query) by setting different mass tolerances associated with different MS methods. The following MS1/MS2 windows were set: MS1-only: 25 ppm/-; Multiple Reaction Monitoring (MRM): 0.7 Da/0.7 Da; and Data-Independent Acquisition (DIA/SWATH): Variable Windows/25 ppm, 25 ppm/25 ppm. The percentage of compounds in the assay library with no interference (the background for each query based on the given parameters), known as the percentage of unique compounds (y-axis), is calculated for each method (x-axis) a) for data acquired at collision energy ranges of 20-50 eV b) and 50-80 eV.

### A.1.4 Combining Identification Information Between Collision Energies

Depending on the molecular mass and structure of the metabolites, fragmentation behavior can be different depending on the collision energy used. Multiple collision energies or collision energy ranges may boost library performance[64]. Here, combining the two collision energy ranges (20-50 eV & 50-80 eV) can boost assay library coverage by 11% (Fig. A.3).

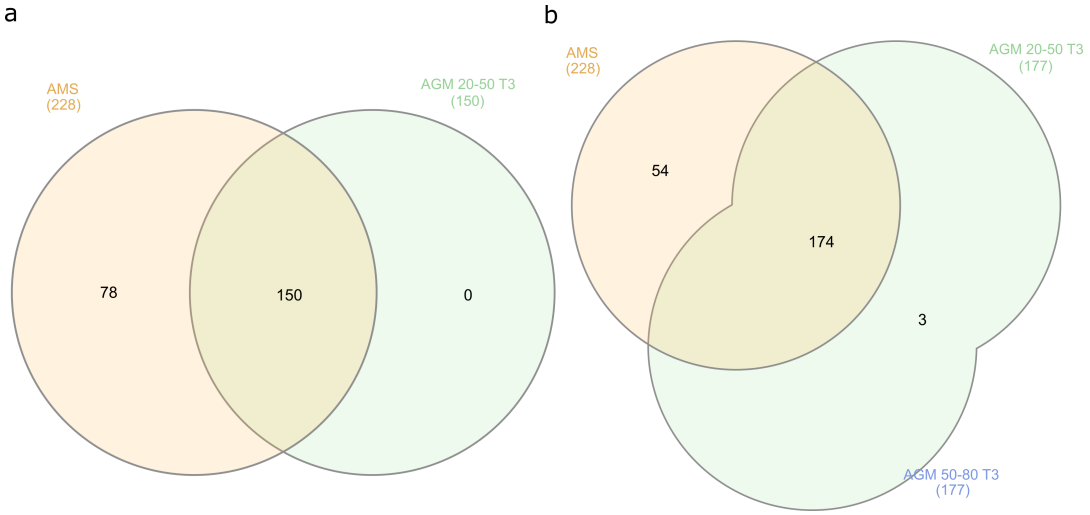


**Figure A.3: Library Coverage** Identification at MS1 level was performed using AccurateMassSearch (AMS). The library with fragment annotation at the MS2 level was generated using 3 transitions (T3) in the AssayGeneratorMetabo (AGM). a) Library generation using 20-50 eV data has a coverage of 66% detected and identified compounds. b) Combining detected and identified compounds from both collision energy ranges (20-50 eV, 50-80 eV) increases the coverage of the target-decoy library to 77%.

### A.1.5 Identification Accuracy for Different Collision Energies

FDR based on the fragmentation tree re-rooting approach has a conservative tendency, with a slight overestimation for data acquired at lower ranges of collision energy (20-50 eV). In comparison, FDR estimates for data acquired at higher ranges of collision energy (50-80 eV) demonstrated an increased abundance of overlapping fragments resulting in more complicated analyses (Fig. A.4). We determined the precision and recall based on different estimated FDR thresholds using the best peak group rank to verify the accuracy of the classifier, which resulted in the majority of all positive results. The area-under-the-curve is 96% and 93%, respectively.

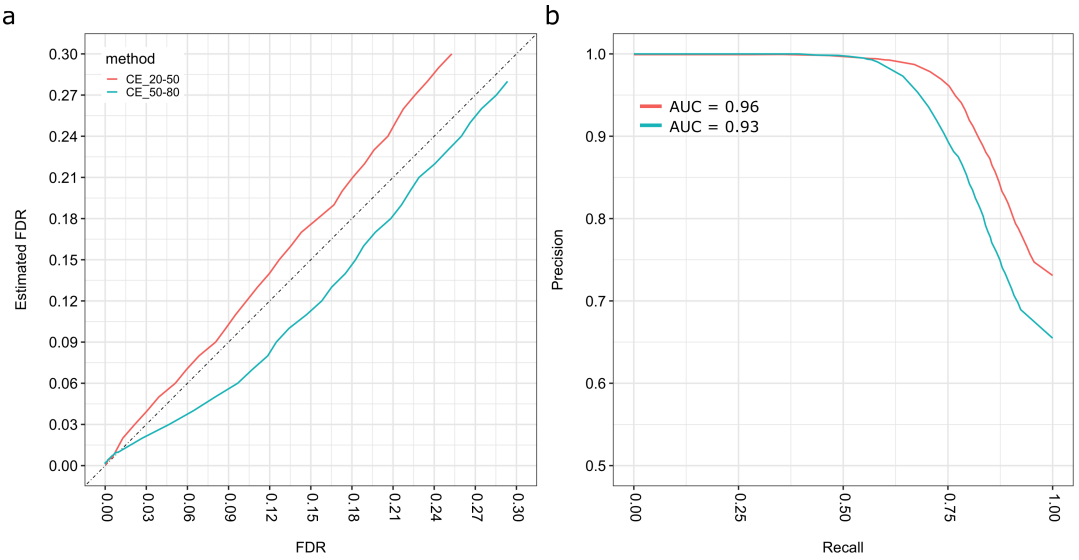


**Figure A.4: Identification Accuracy of DIAMetAlyzer on the Pesticide Spike-In Dataset at Different Collision Energies** a) The estimated FDR versus the pseudo ground truth. The continuous line at 45 degrees shows the optimal values. b) Precision-Recall curve with the area-under-the-curve (AUC (20-50 eV) = 0.96, AUC (50-80 eV) = 0.93).

### A.1.6 Quantification Behavior Collision Energy 50-80 eV

All results were filtered using a 5% FDR threshold. At half of the dilution series (1:1,024), 58 metabolites were detected (Fig. A.5a). The difference in mean standard deviation regarding the theoretical concentration of the automatic and manual analysis as shown in Fig. A.5b. The median coefficient of variation (CV) of quantified signals was below 20% in all technical replicates (Fig. A.5c).

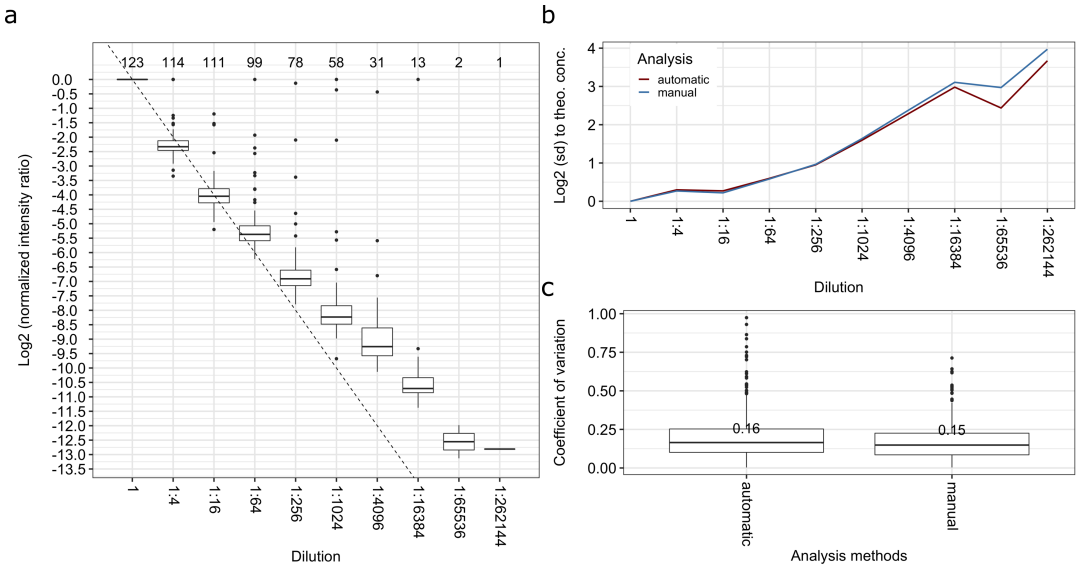


**Figure A.5: Quantification Behavior (CE 50-80 eV)** a) Normalized intensity ratio over the dilution series of data acquired with a collision energy of 50-80 eV (normalized for each metabolite-adduct combination by the intensity of their highest concentration). The dashed line shows the ideal values (fourfold difference to the next dilution). The number on top is the number of metabolites found in the specific dilution at a 5% FDR cutoff. At half of the dilution series (1:1,024) we could detect 58 metabolites. b) Difference in mean standard deviation concerning the theoretical concentration of the automatic and manual analysis. c) Shows the median coefficient of variation (CV across three technical replicates for the automatic and manual analysis). For a and b, only metabolites detected in triplicates and below a 5% FDR threshold were analyzed, and only true positives were considered in the case of panel c.

### A.1.7 Quantification Behavior of Detected Metabolites

Three different quantification behaviors were determined (Fig. A.6). Amidosulfuron (exact mass: 369.04 Da) has a low initial intensity and was detected in the first two dilutions. Azaconazole (exact mass: 229.02 Da), which has a higher measured initial intensity, was detected to a dilution of 1:1,024. Fenpyroximate (exact mass: 421.20 Da), with the highest initial intensity, was detected to a dilution of 1:262,144. Depending on the initial measured intensity, the pesticides could be detected in samples with a higher dilution.

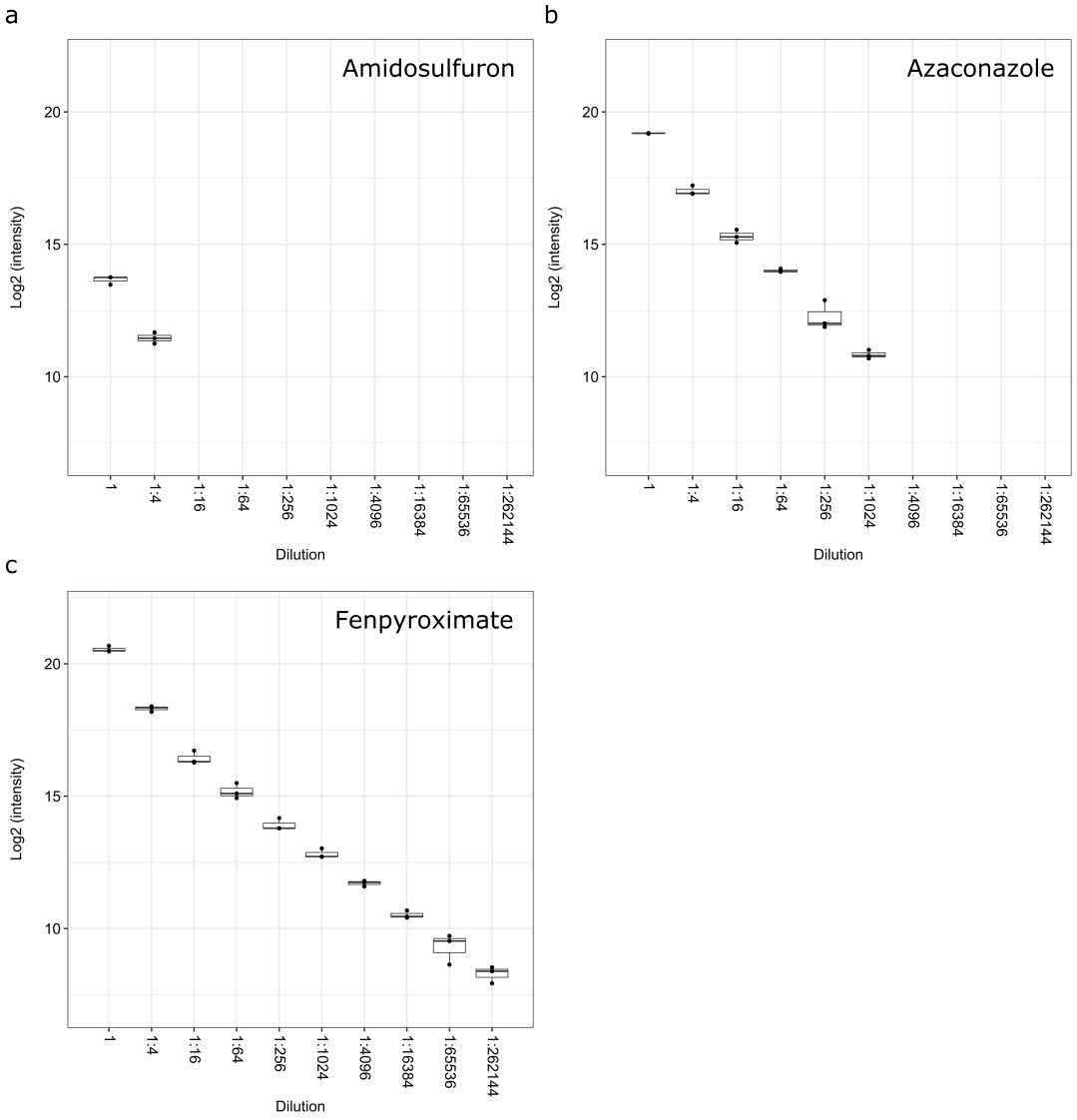


**Figure A.6: Quantification Behaviors over Dilutions for Different Pesticides** a) Amidosulfuron (exact mass: 369.04 Da) has a low initial intensity and was detected in the first two dilutions. b) Azaconazole (exact mass: 229.02 Da) had a higher initial measurable intensity and was detected to a dilution of 1:1,024. c) Fenpyroximate (exact mass: 421.20 Da) was detected to a dilution of 1:262,144 and had the highest initial intensity.

### A.1.8 Pyprophet Model Performance

Performance of the PyProphet[49,50] linear discriminant analysis for metabolomics data. For MS1 and MS2, scoring-only scores that show a low cross-correlation were used (*var_ms1_isotope_overlap_score*, *var_ms1_xcorr_coelution_contrast*, *var_ms1_massdev_score*, *var_ms1_xcorr_coelution*, *var_ms1_isotope_correlation_score*, *var_isotope_overlap_score*, *var_isotope_correlation_score*, *var_intensity_score*, *var_massdev_score*, *var_library_corr*, *var_norm_rt_score*). Using semi-supervised learning, a combined discriminant score was

established. The group discriminant-score (d-score) distribution and density show the decoy and target population results (Fig. A.7a,b). Decoy and false targets are nicely separated from the true targets based on the d-score scoring. The p-value density histogram shows an anti-conservative p-value distribution (Fig. A.7c). For further details regarding changes to PyProphet in terms of metabolomics data processing, please see the source code `https://github.com/PyProphet/pyprophet`

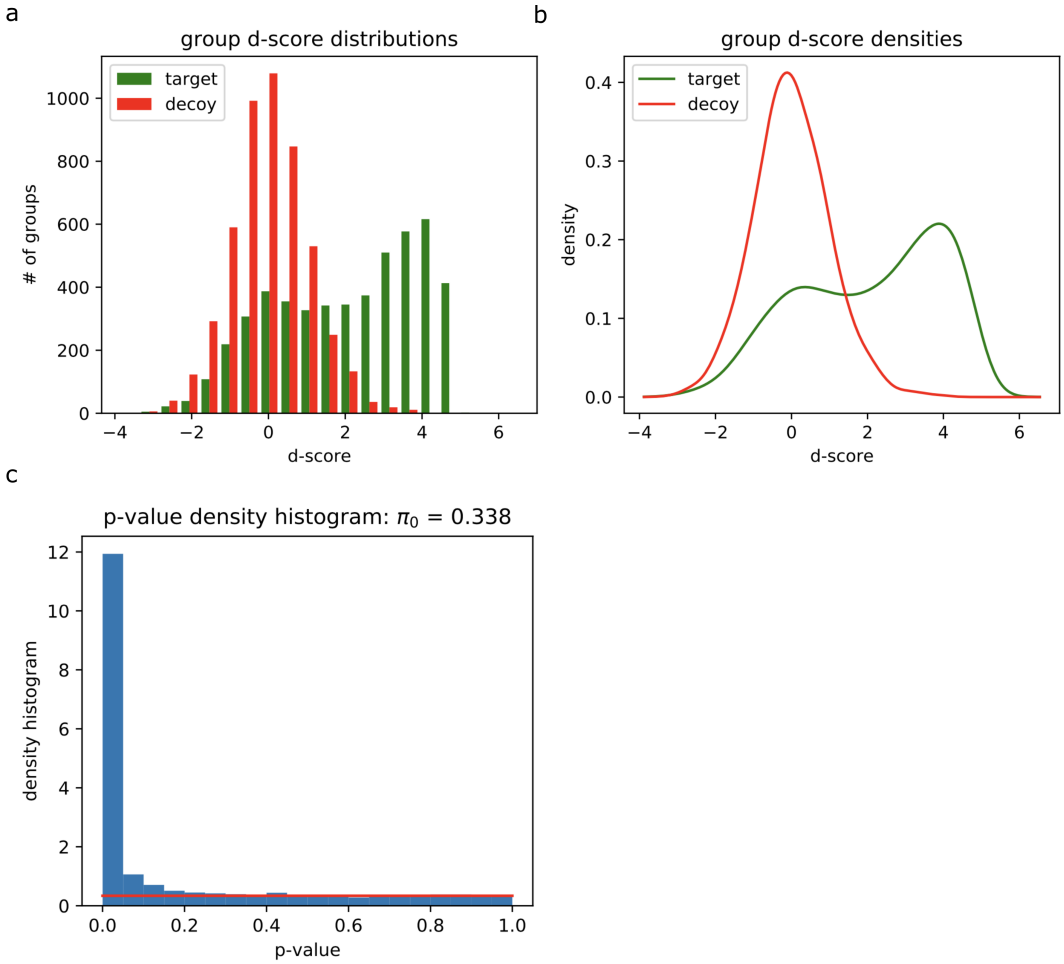


**Figure A.7: PyProphet Target Decoy Evaluation** a,b) The group discriminant-score (d-score) distribution and density show the decoy and target population results. Decoy and false targets are nicely separated from the real targets, based on the d-score scoring. c) The p-value density histogram showing anti-conservative p-value distribution.

### A.1.9 Further Decoy Generation Methods and Evaluation

Various potential decoy methods were evaluated. **Fragdb:** This method performs a random sampling of n fragments with a lower mass than the current precursor. **Linmzperm:** Instead of using masses directly, this method uses the mass difference between the precursor and

fragments. Here, the mass differences were calculated between the precursor and the first fragment, then the first and the second fragment and so on. Afterwards, they were shuffled randomly and subtracted from the precursor or previous fragment, generating new fragment masses. Since the last fragment always has the same total mass difference, the mass of a -$CH_2$ was added to the new fragment mass. If the same mass difference occurred twice, a -$CH_2$ mass was added to the first fragment mass. **Rtperm:** This method performs the same retention time permutation for the precursor and its fragments within a certain minimal retention time difference. This difference should be higher than the retention time parameter set in OpenSWATH and should be in the retention time range used in the experiment. The decoy retention time is generated such as decoy_rt = decoy_rt + ([0.0 - 1.0] - 0.5) * 400, where decoy_rt is initialized with the same value as the target retention time. **Re-rooting:** Here, the fragmentation tree-based method from Passatutto was used. The fragmentation trees were acquired using the fragment annotation via SIRIUS4. The SIRIUS4 format had to be parsed into a Passatutto-compatible format. Then the method was called, and decoy spectra were used to extract transitions. Afterwards, the *n* top intensity peaks were extracted to use in the target-decoy assay library. The following analysis was performed using the 20-50 eV collision energy data. With the introduced "metabolomics" score filter in PyProphet based on scores with a low cross-correlation, all decoy methods perform rather well (Fig. A.8). The re-rooting method has a conservative tendency (slightly overestimating the FDR). We aimed to establish a method that can be used on a multitude of different data sets. In our humble opinion, it is better to stay on the conservative side for FDR filtering, especially if the area under-the-curve for most decoy methods is reported to be in the same range.

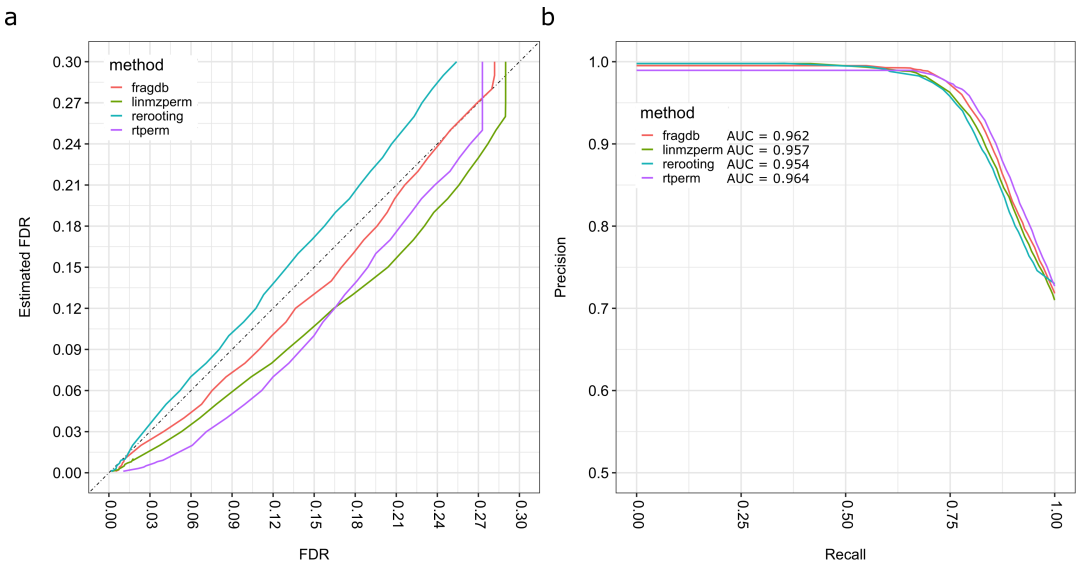


**Figure A.8: Identification Accuracy of DIAMetAlyzer on the Pesticide Spike-In Dataset using Different Decoy Methods** a) The estimated FDR versus the pseudo ground truth. The continuous line at 45 degrees shows the optimal values. b) Precision-Recall curve with the area-under-the-curve (AUC (fragdb) = 0.962 , AUC (linmzperm) = 0.957, AUC (rerooting) = 0.954, AUC (rtperm) = 0.964).

In addition, we looked at the impact of adding a -$CH_2$ mass at different points in the decoy generation method. Adding a -$CH_2$ mass is plausible for organic compounds since it can be added to the carbon backbone. There are two cases where a -$CH_2$ mass can be added to one or more decoy transitions. First, in the case of overlapping target- and decoy-transition masses after extraction, a -$CH_2$ mass was added to the overlapping decoy transition (insertion - in). Second, if re-rooting (Passatutto) of the fragmentation tree was not possible, -$CH_2$ mass was added to all the target transitions, which were then used as decoy transitions (shift). These fallbacks were used in 13% (insertion) and 5% (shift) of the cases. Adding a -$CH_2$ mass in these cases allows the method to have a similar number of targets and decoys, which is not ultimately necessary for the LDA approach used by PyProphet. Here, we tested all combinations of this method. **Original:** Re-rooting with insertion and shift. **WInwoShift:** Re-rooting with insertion without shift. **WoInwoShift:** Re-rooting without insertion and without shift. **WoInwShift:** Re-rooting without insertion with a shift. There are no important differences between the described methods in terms of FDR estimation and precision-recall (Fig. A.9). A major difference in the disparity between the number of decoys for the specific sets (Number of decoys: original = 165, wInwoShift = 155, woInwoShift = 132, woInwShift = 142). As stated previously, a slight discrepancy between the targets and decoys is not problematic for the LDA approach. We decided to further use the re-rooting method with insertion and shift to allow for a robust method applicable on diverse data sets. If the number of decoys is not far below the number of targets, the other methods can be used similarly.

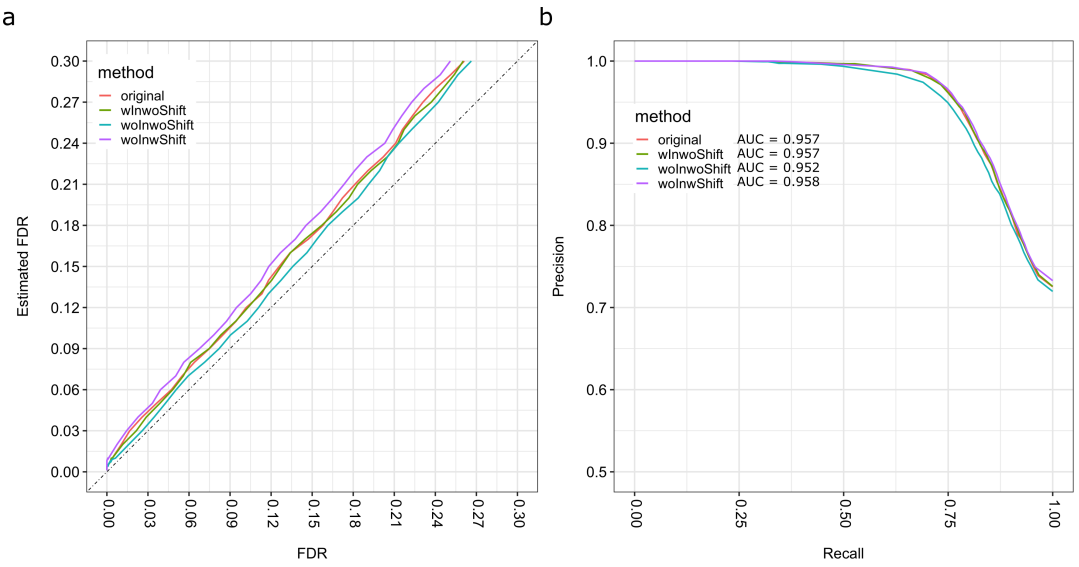


**Figure A.9: Identification Accuracy of DIAMetAlyzer on the Pesticide Spike-In Dataset using the -$CH_2$ Decoy Fallback** a) The estimated FDR versus the pseudo ground truth. The continuous line at 45 degrees shows the optimal values. b) Precision-Recall curve with the area-under-the-curve (AUC (original) = 0.957, AUC (wInwoShift) = 0.957, AUC (woInwoShift) = 0.952, AUC (woInwShift) = 0.958).

### A.1.10 Additional Methods for the Comparison with MetaboDIA

Raw data of MTBLS417 (`https://www.ebi.ac.uk/metabolights/MTBLS417`) was provided by the authors and data was processed as previously described[60]. DDA data was centroided using qtofpeakpicker and msconvert[61] for conversion into mzML/mzXML. DIA data conversion was performed using msconvert. MetaboDIA (Version 1.3) processes DDA and DIA data in centroided mode (mzXML), in contrast to DIAMetAlyzer, which uses centroided DDA (mzML) and profile DIA (mzML) data. An accurate mass search database consisting of HMDB (4.0) and LIPIDMAPS (092020) was constructed, with the entries from the LIPIDMAPS structure file converted into a readable format (Lipid_Enrichtment_prepare_AMS.knwf). Afterwards, duplicates were filtered from the database (reoderDuplicatesMapping.py), and the HMDB and LIPIDMAPS entries were combined (AppendIdentifiersFromOtherDB.py). For MetaboDIA, both databases were combined in a Database.txt file. MetaboDIA was used for the analysis, and in the case of DIAMetAlyzer, an OpenMS development version (14f627e) was used with support for SIRIUS 4.5, allowing internal decoy generation and feature linking for targeted and untargeted experiments using the AssayGeneratorMetabo library generation. For MetaboDIA, the DIA data needed to be processed with DIAUmpire[67] (diaumpire_slurm_mtbls417.sh, params_diaumpire_mtbls417.se_params). DDA and DIA data were processed with XCMS[68] and CAMERA[69] for feature detection and identification. The DDA/DIA workflow from MetaboDIA can be run by specifying the related files, database and adducts (metaboDIA_run.Rmd).

Feature detection, adduct grouping and precursor correction for DIAMetAlyzer were performed using KNIME (20200922_processing_MetaboDIA_FFM_MAD_HRPMC.knwf) followed by accurate mass search (AccurateMassSearch.sh). The assay generation and targeted extraction step were performed on a cluster infrastructure to decrease data processing runtime. The library was constructed by either using prior MS1 identification (agm_mt1_02.sh) or with the addition of unidentified features (agm_mt1_unknown_02.sh). The spectral library stemming from MetaboDIA was converted into an assay library (convertSpetralLibrarytoAssayLibrary_1.4.py) and decoys were generated using the DecoyGeneratorMetabo fragdb method (generateMultiDecoys.sh). A combined library was constructed using DIAMetAlyzer and the converted MetaboDIA library (construct_combined_library.ipynb). All libraries were used for targeted extraction similarly (osw_mt1_67_02.sh). Furthermore, statistical validation was performed using PyProphet (merge_score_export_pyprophet.sh). Results were then processed by DIAlignR[70], which was extended to allow for metabolomics data processing, to reassess values based on chromatogram retention time alignment (DIAlignR.Rmd). Features with an FDR of 0.05 and peak group level 1 were then used for post-processing analysis, which includes library comparison based on the molecular formula, adduct and retention time of a feature (comparison_lib.ipynb). The data was quantitatively compared based on the molecular formula and adduct (compare_lib_and_quant.ipynb), with further comparison of the quantification matrix between features found in MetaboDIA and DIAMetAlyzer to assess the difference of non-targeted vs targeted extraction/quantification (compare_analysis.ipynb). Assessment of differences between the groups was done using limma with Benjamini & Hochberg correction[59] for multiple testing (quant_analysis.Rmd). A part of the identifications of the top significant features was assessed using MASST Search[37].

### A.1.11 The Difference in DDA Feature Detection and Linking of MetaboDIA and DIAMetAlyzer

We are using the DDA data to generate our targeted assay library. Two important steps in this process are feature detection and feature linking. We used the OpenMS feature detection algorithm (FeatureFinderMetabo) developed by Kenar *et al.*, 2014. A short description of the differences of the feature detection algorithms (XCMS[68,69], MzMine[120], Maven[121]) can be found in the publication. In short, FeatureFinderMetabo was compared to XCMS using a simulated ground-truth data set. Here, a similar feature overlap of around 66% was reported, but a higher recall, precision, and F-score for the FeatureFinderMetabo. To compare MetaboDIA and DIAMetAlyzer, we used detected features with a minimum of one mass trace (monoisotopic peak) and linked in at least 50% of the samples. Based on the assay library comparison, DIAMetAlyzer and MetaboDIA overlap in 113 features, whereas MetaboDIA (XCMS/CAMERA) and DIAMetAlyzer (FeatureFinderMetabo) found 39 and 158 exclusive features

respectively. To compare the feature detection and linking based on the library comparison, overlapping and non-overlapping features were exported to an excel sheet, manually converted to EDTA and automatically converted to featureXML. The featureXML was then used for visualization in OpenMS TOPPView. Two DDA files were chosen randomly (one of each sample group: PH697172_pos_IDA-PH697172; CS56422_pos_IDA-CS56422), and the features were manually inspected. Investigating the feature detection and feature linking results of the non-overlapping features in more detail, only 8 of the 39 features in the MetaboDIA library remained undetected by DIAMetAlyzer (FeatureFinderMetabo). The other 31 features are filtered in the assay library generation, discussed in more detail below. In terms of feature detection, XCMS/CAMERA was in some cases not able to separate features (Fig. A.10a), but was able to better detect features with a tailing mass trace (Fig. A.10b) or features that start with a maximum (e.g., at the beginning of the gradient) (Fig. A.10 c,d).

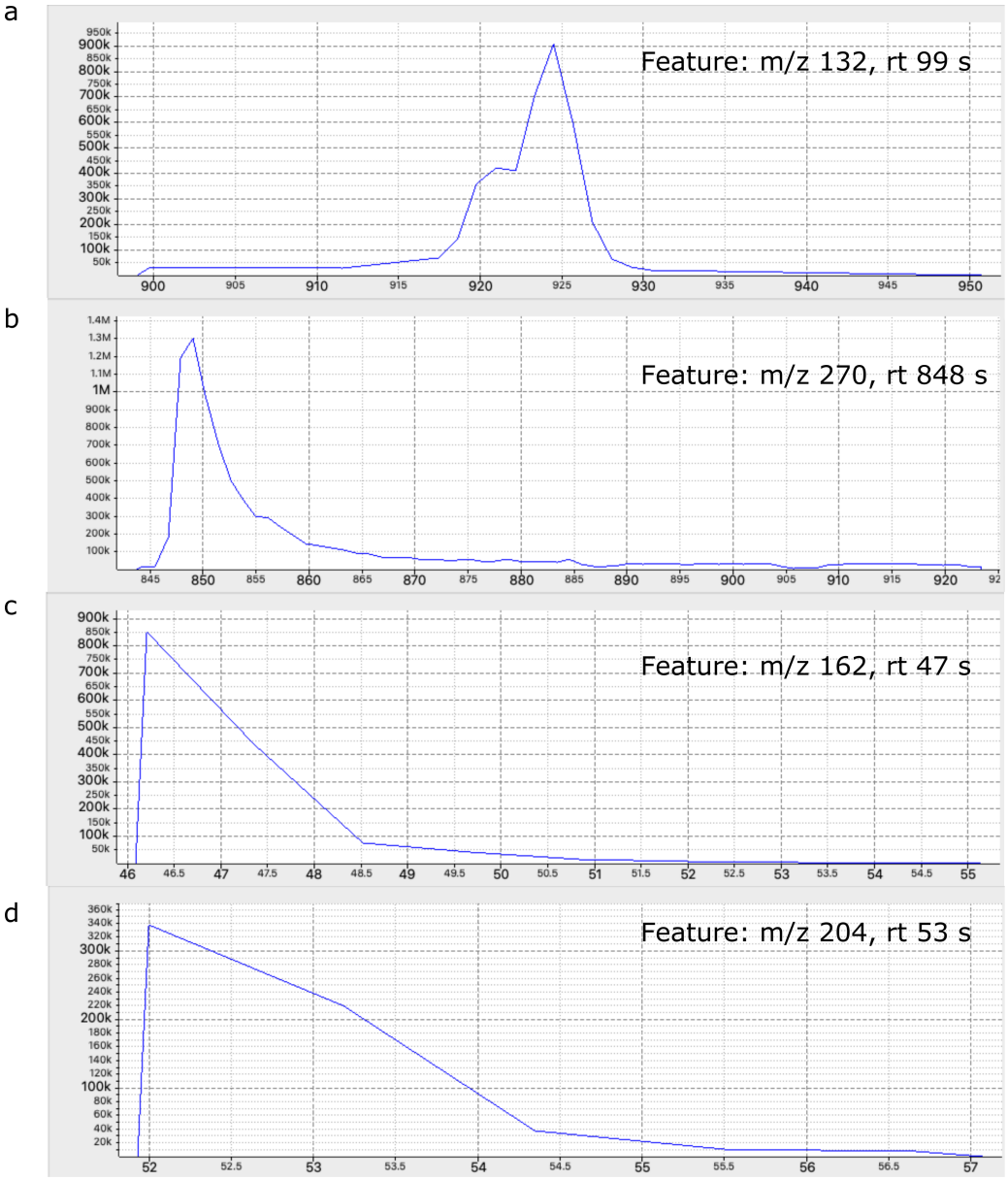


**Figure A.10: Examples for Exclusive Features of the MetaboDIA DDA Feature Detection via XCMS/CAMERA** a) Overlapping features were detected as one feature (m/z 132, rt 99 s). b) Features with tailing mass traces were detected more frequently (m/z 270, rt 99 s). c,d) Features starting at a maximum in the beginning of the gradient were detected more often (m/z 162, rt 47 s; m/z 204, rt 53 s). M/z and retention time (rt) in seconds (s) are associated with the features based on the consensus found over multiple samples.

DIAMetAlyzer (FeatureFinderMetabo) picks up low intensity features more consistently (Fig. A.11a-c). In addition, the manual inspection suggests that FeatureFinderMetabo was able to detect and deconvolute multiple features "hidden" in longer mass traces (Fig. A.11d.)

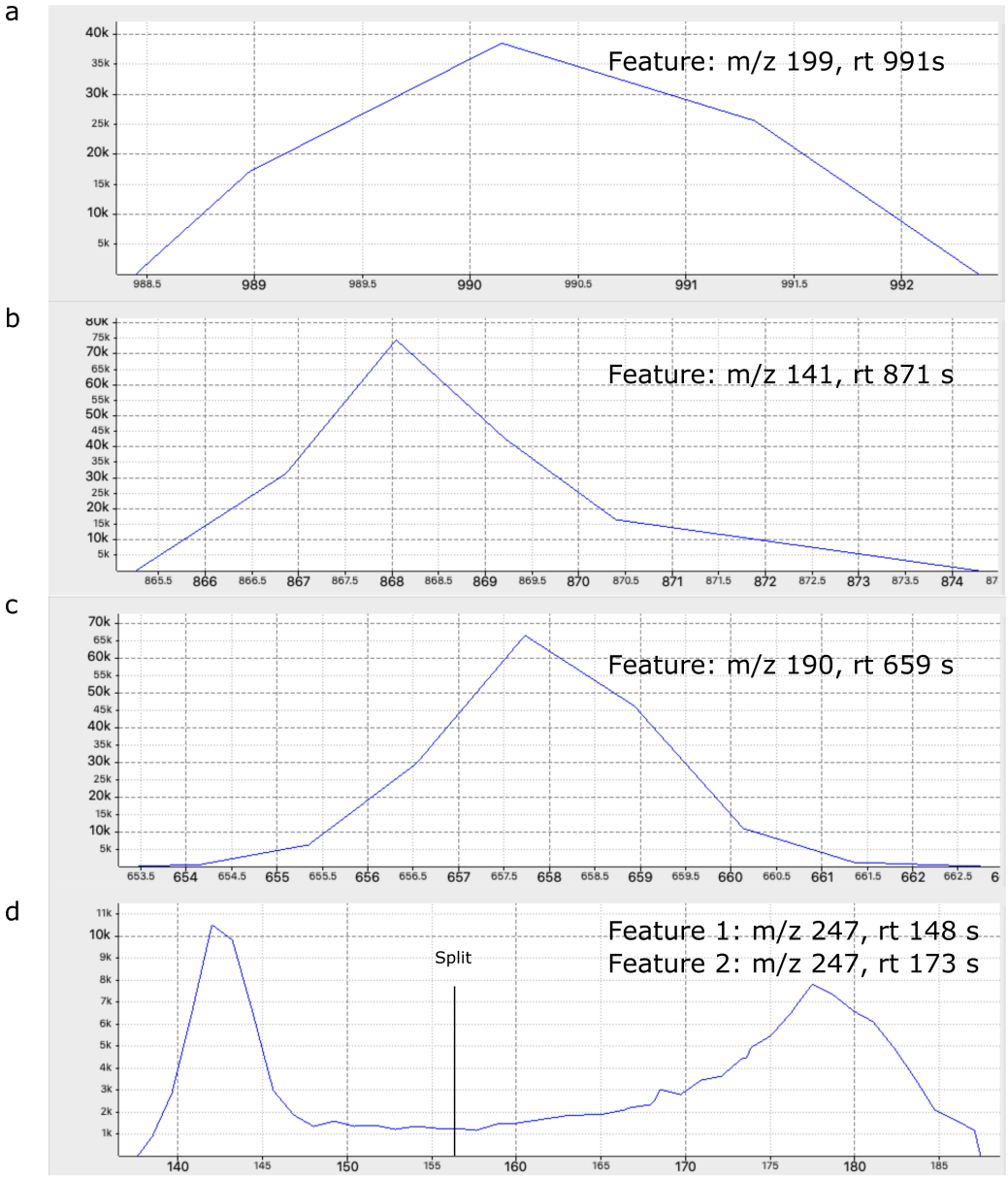


**Figure A.11: Examples for Exclusive Features of DIAMetAlyzer DDA Feature Detection via FeatureFinderMetabo** a,b,c) Low intensity Features with an intensity of 40,000 - 70,000 were detected more consistently (m/z 199, rt 991 s; m/z 141, rt 871 s; m/z 190, rt 659 s). d) FeatureFinderMetabo was able to detect and deconvolute features that were united in a longer mass trace (m/z 247, rt 173 s; m/z 247, rt 173 s). M/z and retention time (rt) in seconds (s) are associated with the features based on the consensus found over multiple samples.

In general, we think with a more refined parameter optimization both algorithms would still be able to detect additional features in this data set.

The presented features in Figure A.10 and A.11 do not represent the main population of features detected in the analysis. In case of Figure A.10 c,d, the analytes are eluting in the beginning

of the gradient and are somewhat cut-off but are still features which can be detected with the XCMS algorithms. In the case of Figure A.11 a-c, the feature intensity span in the file was roughly from 7,523 to 72,903,272. All features shown have a maximum intensity of 75,000 and can be deemed low intensity. The features presented in Figure A.10 and A.11 do not show an optimal chromatographic shape with around 6-8 peaks. Fig. A.12 shows the chromatograms two mid range intensity features and one high intensity feature as representatives of the main feature population in the data set. Please be aware that due to different axis ranges in intensity and retention time, the figures are not directly comparable, but should give an indication of the feature population in the sample.

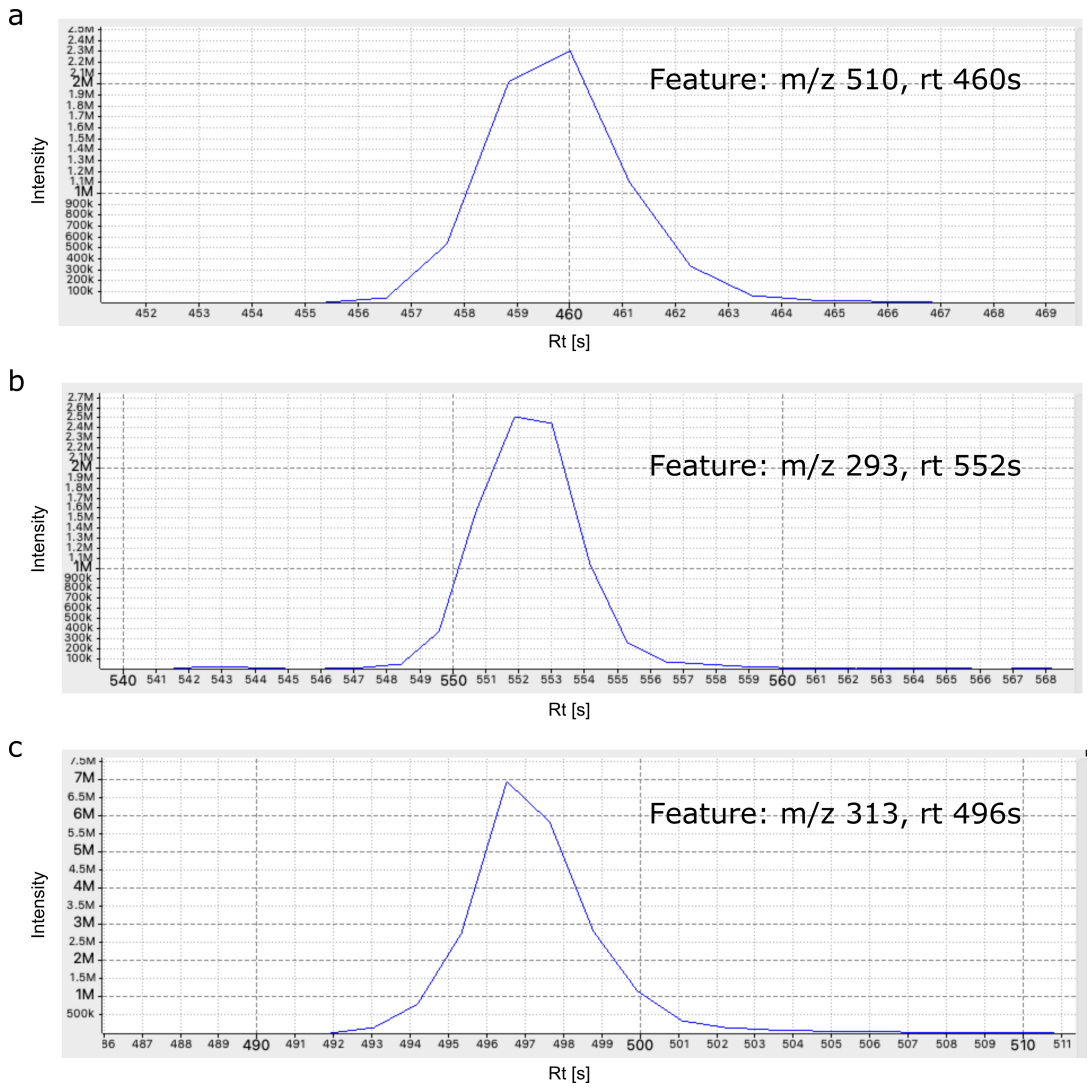


**Figure A.12: Examples for Mid and High Intensity Features Detected in the DDA Data** a,b) Mid intensity features with an intensity of around 2,000,000 were detected in both algorithms. c) High intensity feature with an intensity of 7,000,000. All three features were extracted from DDA data and are shown as representation of the chromatographic outline of the main feature population in the data set.

A big difference between the tools is introduced in the feature linking/assay library generation step. MetaboDIA detects and links features independent of their identification. In DIAMetAlyzer the feature linking is performed after fragment annotation. Here, features with less than four fragment peaks in the MS2 spectra are filtered out by SIRIUS. Further filtering depends on the parameterization of the workflow. Using solely known compounds, features without identification or a correct fragment annotation are filtered out before the linking process. This leads to the filtering of linked features based on the occurrence of their identification. For example, if a feature is linked in 67 samples but is only correctly identified in 13 samples,

in the case of DIAMetAlyzer the feature is filtered out, since it is not available in more than 50% of the samples. This step is performed to ensure a high-quality assay library based on the identification and fragment annotation.
In conclusion, the feature detection from DIAMetAlyzer can detect more features, but the linking is more stringent, dependent on the application, to ensure a high-quality assay library. For a hypothesis-generating approach, we nonetheless would aim for a combined library since validation of the identification can always be performed later. One of the strong points of DIAMetAlyzer is that it allows for the aggregation of information in the assay library e.g., based on the findings of other tools or spectral libraries. In terms of stability and reproducibility, since feature detection, adduct grouping, and identification are distinct steps to construct the assay library, a change in parameters may lead to differences in the results. To improve feature coverage, it is advised to optimize the feature detection parameters to the specific data set. Changing the parameters should not impair the general robustness of the workflow. In terms of reproducibility, KNIME allows to save and export the workflow with the adjusted parameters to fit the experimental setting. We would advise supplying the specific workflow or script which was used for the analysis. This will allow for reproducibility independent of the parameter changes.

### A.1.12 Library Comparison (MetaboDIA vs DIAMetAlyzer)

We compared the different libraries and their entries based on the molecular formula, adduct and retention time information (Fig. A.13). MetaboDIA produces the smallest library, followed by native DIAMetAlyzer. Combining both libraries leads to an increase in the number of features and while staying on the conservative side, by allowing only features with prior identification. Allowing features without MS1 identification in addition to the native DIAMetAlyzer library leads to a further increase of features in the library. Here, SIRIUS is used for the assessment of the molecular formula and the fragment annotation internally.

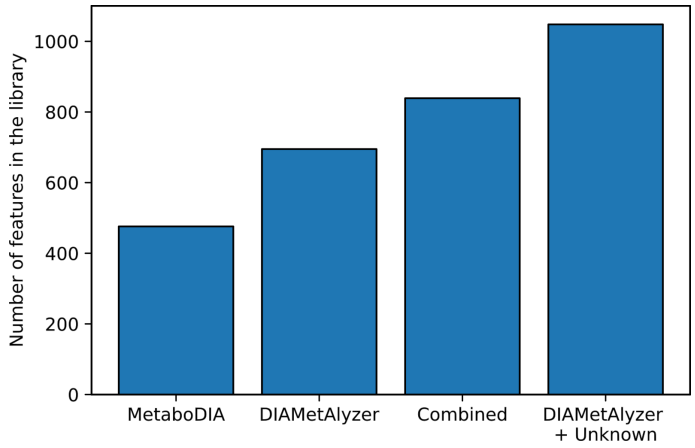


**Figure A.13: Library Comparison MetaboDIA vs DIAMetAlyzer** Comparison of libraries and their entries based on molecular formula, adduct and retention time information. MetaboDIA produces the smallest library (n = 440), followed by native DIAMetAlyzer (n = 695). Combining both libraries leads to an increase of features and stays on the conservative side, by allowing only features with prior identification (n = 839). The option which presents the highest number of entries in the library is to allow features, without an MS1 identification and let SIRIUS assess the molecular formula and fragment annotation (n = 1048).

### A.1.13 Evaluation of the Identification Performance

The top 20 features with the highest significant differences between groups based on DIAMetAlyzer + Unknown quantification were used to assess the identification performance. Within DIAMetAlyzer the molecular formula for features without prior MS1 identification was determined via SIRIUS. In most of these cases, no MS2 spectral library match was detected in the GNPS database using the MASST Search (workflow release version 27), but in almost all cases, a mirror match was detected stemming from another Homo sapiens data set, which suggests a valid feature without spectral library entries. Following such cases could potentially lead to the identification of new compounds which are related to the biological question. Further, three MS1 identifications were directly validated by GNPS spectral library search. Other MS1 identifications could not be validated by GNPS, either since no matching MS2 spectrum could be found or MS2 spectral library search found other possible identification based on the spectrum with a cosine similarity between 0.72 - 0.92. As in most metabolomics studies, the identification of compounds and their biological influence should be re-evaluated in follow up studies.

## Table A.3: Identification of the Top 20 Features Based on DIAMetAlyzer + Unknown

| Features | Identification (MS1 or *de-novo)* | GNPS (spectral library search) | Link to job |
|---|---|---|---|
| 1292_C34H51N11O11_[M+H]+ | UNKNOWN_595 | NoHitsinMS2(GNPS),butcommunity matches | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e81cc1216b5947cdb16dfb5a7f1a36e6 |
| 1376_C22H26O6_[M+H]+ | Porson,IsogingerenoneB,GingerenoneB,?,? | C17H22O10,1-O-Sinapoyl-beta-D-glucose,0.84cosinesimilarity | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2c21a2d7b9a64890864450b47bc403c9 |
| 1385_C22H29NO6_[M+H]+ | UNKNOWN_619 | NoHitsinMS2(GNPS),butcommunity matches | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3731682576c34bcdb2dcdbad71bdf45a |
| 1118_C23H37O7P_[M+H]+ | 1-(5Z,8Z,11Z,14Z,17Z-eicosapentaenoyl)-glycero-3-phosphate,LysoPA(20:5(5Z,8Z,11Z,14Z,17Z)/0:0),1-(5Z,8Z,11Z,14Z,17Z-eicosapentaenoyl)-glycero-3-phosphate | NoHitsinMS2(GNPS) | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0222cc9ff8db45aa8ff53a1c7af64a17 |
| 382_C21H28N6O4S_[M+H]+ | N-Desmethylsildenafil(UK-103,320) | NoHitsinMS2(GNPS) | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7be87fce1597468591538a00ef4ab352 |
| 228_C16H19N3O5_[M+H]+ | Glutamyltryptophan,Tryptophyl-Glutamate,gamma-Glutamyltryptophan | SpectralMatchtoTrp-GlufromNIST14 | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=85ee7e43284a43a2ab7717cc52d7d3ff |
| 874_C38H42O4_[M+K]+ | UNKNOWN_564 | C33H36N4O6-Bilirubin,0.82cosinesimilarity | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=fd36000610ce4cec8301d4ee26449458 |
| 1482_C12H18O4S2_[M+H]+ | Isoprothiolane | NoHitsinMS2(GNPS) | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3f281e9239cf410eaaff373fa1751deb |
| 532_C18H23N3O6_[M+H]+ | Imidaprilat | NoHitsinMS2(GNPS) | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7e04a1df3e3d440e959207897581326a |
| 502_C14H22ClN3O2_[M+H]+ | Metoclopramide | Massbank,Metoclopramide\|4-Amino-5-chloro-N-\[2-(diethylamino)ethyl]-2-methoxybenzamide | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=11cc03314f344b4a9e5b51efdfdae050 |
| 1234_C30H42O12_[M+Na]+ | 3-O(beta-D-glucopyranosyl)-3beta,5beta,14beta,16beta-tetrahydroxy\19-oxo-bufa-20,22-dienolide | 3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonatefromNIST14,0.73consinesimilarity | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a4d585bc4ff44b0c8f583836c8251f33 |
| 1594_C28H50NO8P_[M+H]+ | 2-(11R-hydroxy-5Z,8Z,12E,14Z-eicosatetraenoyl)-sn-glycero-3-phosphocholine,2-(15R-hydroxy-5Z,8Z,12E,14Z-eicosatetraenoyl)-sn-glycero-3-phosphocholine,2-(15S-hydroxy-5Z,8Z,12E,14Z-eicosatetraenoyl)-sn-glycero-3-phosphocholine | NoHitsinMS2(GNPS),butcommunity matches | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=193fa5748add45c981845cc037fc27a8 |
| 598_C14H16N2O4_[M+H]+ | N-lactoyl-Tryptophan | NoHitsinMS2(GNPS),butcommunity matches | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=87b4f973666f46af82d060ccc4985018 |

| | | | |
|---|---|---|---|
| 1524_C17H37NO2_[M+H]+ | heptadecasphinganine | C16H33NO3,Lauryldiethanolamide,0.92cosinesimilarity | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=66031f129198 |
| 280_C14H18N2O5_[M+H]+ | gamma-Glutamylphenylalanine,Aspartame,Hydroxyprolyl-Tyrosine,Phenylalanyl-Glutamate,Tyrosyl-Hydroxyproline,Glutamylphenylalanine,4'-tert-Butyl-2',6'-dimethyl-3',5'-dinitroacetophenone,L-gamma-Glutamyl-beta-\phenyl-beta-L-alanine | NoHitsinMS2(GNPS),butcommunity matches | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4e5ab2de57b8a04d4ff3 |
| 1073_C17H36O7_[M+H]+ | UNKNOWN_500 | NoHitsinMS2(GNPS) | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=610b621bf01340e6bd328e148935fe75 |
| 472_C26H29N3O6_[M+H]+ | Nicardipine | NoHitsinMS2(GNPS),butcommunity matches | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2ab6dcc7f0184f7d97a13d71051f63c9 |
| 995_C20H39N5O4_[M+H]+ | UNKNOWN_476 | NoHitsinMS2(GNPS),butcommunity matches | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=daa5c6acdf6d490385241281af97b9b4 |
| 1552_C26H5ON08P_[M+H]+ | 1-(9Z-hexadecenoyl)-2-acetyl-sn-glycero-3\-phosphocholine | C27H55N1O7P1,0.72cosinesimiliarity | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8f1faf40c6a9400d8d536687ff896d51 |
| 1970_C26H54NO6P_[M+H]+ | LysoPC(P-18:0),1-(11Z-octadecenyl)-sn-glycero-3-phosphocholine,1-(9Z-octadecenyl)-sn-glycero-3-phosphocholine,1-(1Z-octadecenyl)-sn-glycero-3-phosphocholine | 1-(1Z-Octadecenyl)-sn-glycero-3-phosphocholine\fromNIST14,0.92cosinesimilarity | https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7dfc1cadabc742508458c64614953217 |

### A.1.14 Comparison of DDA and DIA Data

We visually compared the DDA and DIA XICs for the highest concentration and chose two representatives (Fig. A.14). Both show a similar chromatogram shape on the MS1 level. At the MS2 level, the peak depends on the trigger time of the instrument (the chromatogram shape is based on the interpolation used in the Skyline visualization). For DIA MS2 the chromatogram is clearly visible and co-elutes exactly with the MS1 chromatogram.

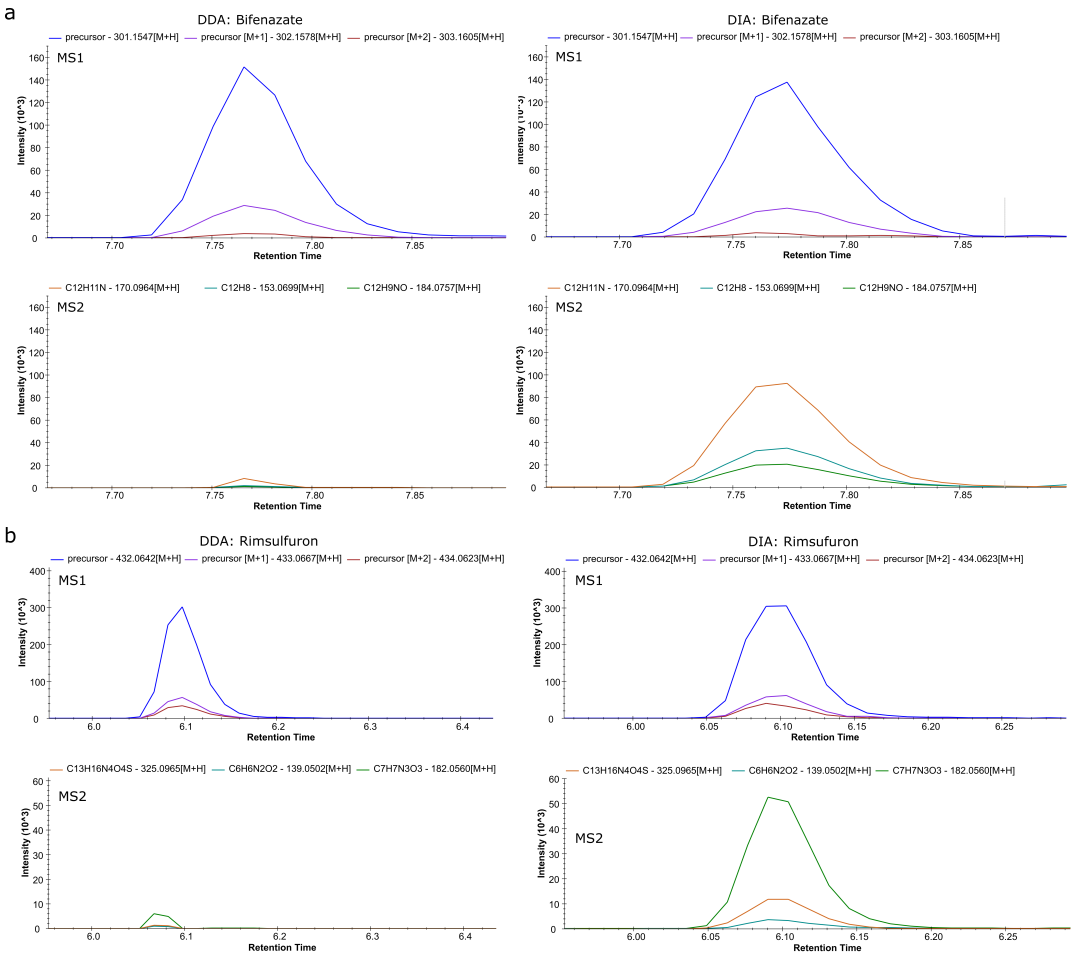


**Figure A.14: Examples for MS1 and MS2 XICs from DDA and DIA Data** The study data representative pesticides Bifenazate (a) and Rimsulfuron (b) show a similar MS1 XIC in DDA and DIA.

### A.1.15 Limit of Detection

We calculated the limit of detection (LOD), using the unfiltered results and the definition of LOD = S/N > 10 for each pesticide (Table A.4). We used the intensity at the lowest dilution the compound was still detected via targeted extraction as noise. From there we calculated the concentration based on the exact mass, the initial concentration of 1 ng/µl and

the corresponding dilution of the compound. Please see *lod.RMD* at `https://github.com/KohlbacherLab/DIAMetAlyzer_additional_code` for details.

**Table A.4: Limit of Detection**

| compound name | molecular formular | last SN over threshold | last dilution over threshold | LOD [fmol/µl] |
|---|---|---|---|---|
| Acephate | $C_4H_{10}NO_3PS$ | 15 | 3 | 342 |
| Sulfadiazole | $C_7H_{12}N_4O_3S_2$ | 15 | 7 | 237 |
| Carboxin | $C_{12}H_{13}NO_2S$ | 11 | 7 | 266 |
| Flusilazole | $C_{16}H_{15}F_2N_3Si$ | 20 | 5 | 198 |
| Phosmet | $C_{11}H_{12}NO_4PS_2$ | 44 | 4 | 197 |
| Thifensulfurson methyl | $C_{12}H_{13}N_5O_6S_2$ | 23 | 6 | 161 |
| Desmedipham | $C_{16}H_{16}N_2O_4$ | 14 | 7 | 208 |
| Malaoxon | $C_{10}H_{19}O_7PS$ | 17 | 6 | 199 |
| Metribuzin | $C_8H_{14}N_4OS$ | 14 | 6 | 292 |
| Nitenpyram | $C_{11}H_{15}ClN_4O_2$ | 14 | 4 | 231 |
| Dimethomorph | $C_{21}H_{22}ClNO_4$ | 14 | 6 | 161 |
| Fenamiphos | $C_{13}H_{22}NO_3PS$ | 22 | 2 | 206 |
| Metsulfuron-methyl | $C_{14}H_{15}N_5O_6S$ | 13 | 6 | 164 |
| Trietazin | $C_9H_{16}ClN_5$ | 19 | 7 | 273 |
| Halofenozide | $C_{18}H_{19}ClN_2O_2$ | 11 | 5 | 189 |
| Myclobutanil | $C_{15}H_{17}ClN_4$ | 29 | 5 | 217 |
| Halosulfuron-methyl | $C_{13}H_{15}ClN_6O_7S$ | 15 | 4 | 144 |
| Hexaconazole | $C_{14}H_{17}Cl_2N_3O$ | 20 | 7 | 200 |
| Iprovalicarb | $C_{18}H_{28}N_2O_3$ | 14 | 6 | 195 |
| Triasulfuron | $C_{14}H_{16}ClN_5O_5S$ | 13 | 7 | 156 |
| Azoxystrobin | $C_{22}H_{17}N_3O_5$ | 24 | 6 | 155 |
| Bifenazate | $C_{17}H_{20}N_2O_3$ | 19 | 3 | 208 |
| Fluopicolide | $C_{14}H_8Cl_3F_3N_2O$ | 21 | 8 | 164 |
| Metconazole | $C_{17}H_{22}ClN_3O$ | 24 | 6 | 196 |
| Rimsulfuron | $C_{14}H_{17}N_5O_7S_2$ | 13 | 6 | 145 |
| Zoxamide | $C_{14}H_{16}Cl_3NO_2$ | 15 | 6 | 187 |
| Bifenazate | $C_{17}H_{20}N_2O_3$ | 79 | 9 | 208 |
| Clethodim | $C_{17}H_{26}ClNO_3S$ | 14 | 2 | 174 |
| Isoprothiolane | $C_{12}H_{18}O_4S_2$ | 27 | 8 | 215 |
| Methoprotryne | $C_{11}H_{21}N_5OS$ | 15 | 7 | 231 |
| Fenamidone | $C_{17}H_{17}N_3OS$ | 12 | 5 | 201 |
| Fluoxastrobin | $C_{21}H_{16}ClFN_4O_5$ | 12 | 8 | 136 |
| Isoprothiolane | $C_{12}H_{18}O_4S_2$ | 23 | 7 | 215 |
| Isoxaflutole | $C_{15}H_{12}F_3NO_4S$ | 12 | 3 | 174 |
| Propioconazole | $C_{15}H_{17}Cl_2N_3O_2$ | 11 | 7 | 183 |
| Secbumeton | $C_{10}H_{19}N_5O$ | 11 | 7 | 278 |
| Amidosulfuron | $C_9H_{15}N_5O_7S_2$ | 12 | 7 | 169 |
| Bispyribac sodium salt | $C_{19}H_{18}N_4O_8$ | 12 | 6 | 138 |
| Flufenacet | $C_{14}H_{13}F_4N_3O_2S$ | 11 | 8 | 172 |
| Fluoxastrobin | $C_{21}H_{16}ClFN_4O_5$ | 30 | 1 | 136 |
| Triadimefon | $C_{14}H_{16}ClN_3O_2$ | 17 | 5 | 213 |
| Azinphos-Ethyl | $C_{12}H_{16}N_3O_3PS_2$ | 10 | 4 | 181 |
| Chlorfenvinphos | $C_{12}H_{14}Cl_3O_4P$ | 21 | 7 | 175 |
| Cyprodinil | $C_{14}H_{15}N_3$ | 11 | 6 | 278 |
| Furathiocarb | $C_{18}H_{26}N_2O_5S$ | 21 | 7 | 164 |
| Methiocarb | $C_{11}H_{15}NO_2S$ | 24 | 3 | 278 |
| Methoxyfenozide | $C_{22}H_{28}N_2O_3$ | 19 | 3 | 170 |
| Tribenuron methyl | $C_{15}H_{17}N_5O_6S$ | 25 | 4 | 158 |
| Aminocarb | $C_{11}H_{16}N_2O_2$ | 80 | 9 | 300 |
| Lenacil | $C_{13}H_{18}N_2O_2$ | 28 | 1 | 267 |
| Tricyclazol | $C_9H_7N_3S$ | 12 | 6 | 331 |
| Dimoxystrobin | $C_{19}H_{22}N_2O_3$ | 14 | 7 | 192 |
| Flazasulfuron | $C_{13}H_{12}F_3N_5O_5S$ | 14 | 6 | 154 |
| Ipconazole | $C_{18}H_{24}ClN_3O$ | 17 | 5 | 188 |
| Methoxyfenozide | $C_{22}H_{28}N_2O_3$ | 16 | 3 | 170 |
| Propyzamide | $C_{12}H_{11}Cl_2NO$ | 13 | 6 | 245 |
| Spiromesifen | $C_{23}H_{30}O_4$ | 12 | 6 | 169 |
| Isofenphos-methyl | $C_{14}H_{22}NO_4PS$ | 17 | 9 | 189 |
| Prometon | $C_{10}H_{19}N_5O$ | 24 | 5 | 278 |
| Tepraloxydim | $C_{17}H_{24}ClNO_4$ | 10 | 3 | 183 |
| Alanycarb | $C_{17}H_{25}N_3O_4S_2$ | 22 | 7 | 157 |
| Fenhexamid | $C_{14}H_{17}Cl_2NO_2$ | 14 | 5 | 208 |
| Coumaphos | $C_{14}H_{16}ClO_5PS$ | 19 | 7 | 173 |
| Diflufenican | $C_{19}H_{11}F_5N_2O_2$ | 26 | 6 | 159 |
| Mepanipyrim | $C_{14}H_{13}N_3$ | 12 | 5 | 280 |
| Phenmediphan | $C_{16}H_{16}N_2O_4$ | 29 | 6 | 208 |

| | | | | |
|---|---|---|---|---|
| Rotenone | $C_{23}H_{22}O_6$ | 21 | 5 | 159 |
| Malathion | $C_{10}H_{19}O_6PS_2$ | 20 | 7 | 189 |
| Bupirimate | $C_{13}H_{24}N_4O_3S$ | 13 | 7 | 198 |
| Hexaflumuron | $C_{16}H_8Cl_2F_6N_2O_3$ | 18 | 6 | 136 |
| Mandipropamid | $C_{23}H_{22}ClNO_4$ | 19 | 7 | 152 |
| Benzoximate | $C_{18}H_{18}ClNO_5$ | 15 | 2 | 172 |
| Dursban | $C_9H_{11}Cl_3NO_3PS$ | 14 | 6 | 179 |
| Mandipropamid | $C_{23}H_{22}ClNO_4$ | 17 | 5 | 152 |
| Picoxystrobin | $C_{18}H_{16}F_3NO_4$ | 19 | 6 | 170 |
| Proquinazid | $C_{14}H_{17}IN_2O_2$ | 10 | 7 | 168 |
| Benzoximate | $C_{18}H_{18}ClNO_5$ | 23 | 2 | 172 |
| Boscalid | $C_{18}H_{12}Cl_2N_2O$ | 11 | 7 | 183 |
| Picoxystrobin | $C_{18}H_{16}F_3NO_4$ | 13 | 3 | 170 |
| Spirodiclofen | $C_{21}H_{24}Cl_2O_4$ | 11 | 6 | 152 |
| Ethofumesate | $C_{13}H_{18}O_5S$ | 27 | 3 | 218 |
| Mecarbam | $C_{10}H_{20}NO_5PS_2$ | 26 | 6 | 190 |
| Pyraclostrobin | $C_{19}H_{18}ClN_3O_4$ | 12 | 6 | 161 |
| Temephos | $C_{16}H_{20}O_6P_2S_3$ | 18 | 6 | 134 |
| Mixture of Avermectin B1a and B1b | $C_{48}H_{72}O_{14}$ | 17 | 2 | 72 |
| Azamethiphos | $C_9H_{10}ClN_2O_5PS$ | 18 | 8 | 193 |
| Carbaryl | $C_{12}H_{11}NO_2$ | 10 | 4 | 311 |
| Carbendazim | $C_9H_9N_3O_2$ | 34 | 7 | 327 |
| Ethyoxyquin | $C_{14}H_{19}NO$ | 18 | 2 | 288 |
| Mexacarbate | $C_{12}H_{18}N_2O_2$ | 23 | 6 | 281 |
| Pirimicarb | $C_{11}H_{18}N_4O_2$ | 18 | 5 | 262 |
| Moxidectin | $C_{37}H_{53}NO_8$ | 15 | 3 | 98 |
| Trifloxystrobin | ceC20H19F3N2O4 | 14 | 7 | 153 |
| Pyripoxyfen | $C_{20}H_{19}NO_3$ | 23 | 4 | 195 |
| Quinalfos | $C_{12}H_{15}N_2O_3PS$ | 15 | 8 | 210 |
| Silthiofam | $C_{13}H_{21}NOSSi$ | 15 | 7 | 234 |
| Difenoconazole | $C_{19}H_{17}Cl_2N_3O_3$ | 52 | 7 | 154 |
| Fipronil | $C_{12}H_4Cl_2F_6N_4OS$ | 15 | 6 | 143 |
| Cyazofamid | $C_{13}H_{13}ClN_4O_2S$ | 14 | 3 | 193 |
| Phenthoate | $C_{12}H_{17}O_4PS_2$ | 15 | 9 | 195 |
| Ethofumesate | $C_{13}H_{18}O_5S$ | 14 | 2 | 218 |
| Fenpyroximate | $C_{24}H_{27}N_3O_4$ | 11 | 7 | 148 |
| Tolylfluanid | $C_{10}H_{13}Cl_2FN_2O_2S_2$ | 37 | 2 | 181 |
| Carfentrazone-ethyl | $C_{15}H_{14}Cl_2F_3N_3O_3$ | 17 | 8 | 152 |
| Tolclosof methyl | $C_9H_{11}Cl_2O_3PS$ | 21 | 6 | 208 |
| Carfentrazone-ethyl | $C_{15}H_{14}Cl_2F_3N_3O_3$ | 37 | 3 | 152 |
| Fenazaquin | $C_{20}H_{22}N_2O$ | 21 | 4 | 204 |
| Indoxacarb | $C_{22}H_{17}ClF_3N_3O_7$ | 19 | 7 | 119 |
| Kresoxim-methyl | $C_{18}H_{19}NO_4$ | 21 | 5 | 200 |
| Indoxacarb | $C_{22}H_{17}ClF_3N_3O_7$ | 33 | 5 | 119 |
| Kresoxim-methyl | $C_{18}H_{19}NO_4$ | 40 | 3 | 200 |
| Pyridaben | $C_{19}H_{25}ClN_2OS$ | 15 | 3 | 172 |
| Profenofos | $C_{11}H_{15}BrClO_3PS$ | 13 | 6 | 168 |
| Triflumuron | $C_{15}H_{10}ClF_3N_2O_3$ | 11 | 7 | 175 |
| Flumetsulam | $C_{12}H_9F_2N_5O_2S$ | 26 | 7 | 192 |
| Fuberidazole | $C_{11}H_8N_2O$ | 16 | 7 | 340 |
| Sulfentrazone | $C_{11}H_{10}Cl_2F_2N_4O_3S$ | 13 | 5 | 162 |
| Thiabendazole | $C_{10}H_7N_3S$ | 19 | 6 | 311 |
| Tebufanpyrad | $C_{18}H_{24}ClN_3O$ | 15 | 2 | 188 |
| Pirmiphos-methyl | $C_{11}H_{20}N_3O_3PS$ | 16 | 7 | 205 |
| Picolinafen | $C_{19}H_{12}F_4N_2O_2$ | 19 | 8 | 166 |
| Quinoxyfen | $C_{15}H_8Cl_2FNO$ | 27 | 7 | 204 |
| Teflubenzuron | $C_{14}H_6Cl_2F_4N_2O_2$ | 12 | 6 | 164 |
| Oxadiazon | $C_{15}H_{18}Cl_2N_2O_3$ | 17 | 7 | 182 |
| Metrafenone | $C_{19}H_{21}BrO_5$ | 25 | 5 | 153 |
| Pendimethalin (Penoxalin) | $C_{13}H_{19}N_3O_4$ | 14 | 6 | 222 |
| Propaquizafop | $C_{22}H_{22}ClN_3O_5$ | 15 | 5 | 141 |
| Benfuracarb | $C_{20}H_{30}N_2O_5S$ | 23 | 1 | 152 |
| Benfuracarb | $C_{20}H_{30}N_2O_5S$ | 11 | 3 | 152 |
| Fosthiazate | $C_9H_{18}NO_3PS_2$ | 15 | 6 | 221 |
| Metamitron | $C_{10}H_{10}N_4O$ | 16 | 3 | 309 |
| Phospamidon (Mix of isomers) | $C_{10}H_{19}ClNO_5P$ | 15 | 7 | 209 |
| Sulfentrazone | $C_{11}H_{10}Cl_2F_2N_4O_3S$ | 11 | 7 | 162 |
| Thiametoxam | $C_8H_{10}ClN_5O_3S$ | 24 | 6 | 215 |
| Flufenoxuron | $C_{21}H_{11}ClF_6N_2O_3$ | 12 | 8 | 128 |
| Hexythiazox | $C_{17}H_{21}ClN_2O_2S$ | 12 | 6 | 178 |
| Carbofuran | $C_{12}H_{15}NO_3$ | 15 | 4 | 283 |
| Chloridazon | $C_{10}H_8ClN_3O$ | 48 | 8 | 283 |

| | | | | |
|---|---|---|---|---|
| Fosthiazate | $C_9H_{18}NO_3PS_2$ | 23 | 7 | 221 |
| Imidacloprid | $C_9H_{10}ClN_5O_2$ | 16 | 6 | 245 |
| Mesosulfuron-methyl | $C_{17}H_{21}N_5O_9S_2$ | 14 | 6 | 124 |
| Triticonazole | $C_{17}H_{20}ClN_3O$ | 32 | 5 | 197 |
| Azaconazole | $C_{12}H_{11}Cl_2N_3O_2$ | 10 | 7 | 209 |
| Cymoxanil | $C_7H_{10}N_4O_3$ | 23 | 2 | 316 |
| Fenarimol | $C_{17}H_{12}Cl_2N_2O$ | 18 | 6 | 189 |
| Fluometuron | $C_{10}H_{11}F_3N_2O$ | 13 | 6 | 269 |
| Nitenpyram | $C_{11}H_{15}ClN_4O_2$ | 11 | 5 | 231 |
| Cymiazol hydrochloride | $C_{12}H_{14}N_2S$ | 12 | 7 | 287 |
| Ethirimol | $C_{11}H_{19}N_3O$ | 12 | 7 | 299 |
| Flumioxazin | $C_{19}H_{15}FN_2O_4$ | 13 | 2 | 177 |
| Fluquinconazole | $C_{16}H_8Cl_2FN_5O$ | 14 | 7 | 167 |
| Flutriafol | $C_{16}H_{13}F_2N_3O$ | 19 | 6 | 208 |
| Diethofencarb | $C_{14}H_{21}NO_4$ | 14 | 7 | 234 |
| Dimethachlor | $C_{13}H_{18}ClNO_2$ | 36 | 7 | 245 |
| Ethoprop | $C_8H_{19}O_2PS_2$ | 18 | 6 | 258 |
| Foramsulfron | $C_{17}H_{20}N_6O_7S$ | 17 | 6 | 138 |
| Metobromuron | $C_9H_{11}BrN_2O_2$ | 16 | 6 | 242 |
| Chlorantraniliprole | $C_{18}H_{14}BrCl_2N_5O_2$ | 11 | 7 | 130 |
| Guthion | $C_{10}H_{12}N_3O_3PS_2$ | 11 | 6 | 197 |
| Methabenzthiazurone | $C_{10}H_{11}N_3OS$ | 10 | 7 | 283 |
| Oxadixyl | $C_{14}H_{18}N_2O_4$ | 13 | 4 | 225 |
| Quinoclamine | $C_{10}H_6ClNO_2$ | 35 | 2 | 302 |
| Spirotetramat | $C_{21}H_{27}NO_5$ | 14 | 7 | 167 |
| Tebuconazole | $C_{16}H_{22}ClN_3O$ | 16 | 6 | 203 |
| Tebuthiuron | $C_9H_{16}N_4OS$ | 12 | 6 | 274 |

## A.2 Reporting Standardization in Metabolomics: MzTab-M

Due to the step-by-step introduction of the identification data structure in OpenMS, we used an identification integrated *AccurateMassSearch*-pipeline as a proof-of-concept for the *MzTab-M* export. We used the the OpenMS Tutorial example data (2012_02_03_PStd_10_1) available at `https://abibuilder.cs.uni-tuebingen.de/archive/openms/Tutorials/Example_Data/Metabolomics/`. Since the MzTab-M export was not merged into the main repository, we used the "idf_oms" branch (`https://github.com/OpenMS/OpenMS/tree/idf_ams`, commit *b555882221*).
Feature detection of the centroided *.mzML* was performed using the *FeatureFinderMetabo*, to detect analytes by m/z, charge and retention time (Listing A.1).

**Listing A.1:** Command: FeatureFinderMetabo

```
1  FeatureFinderMetabo
2  -in 2012_02_03_PStd_10_1.mzML
3  -out 2012_02_03_PStd_10_1.featureXML
```

In the next step, adduct grouping was performed by using the *MetaboliteAdductDecharger*, to assess possible adducts for the detected features (Listing A.2). Here, we allowed the adducts [M+H]+ and [M+Na]+, as well as a minimum/maximum charge of one.

**Listing A.2:** Command: MetaboliteAdductDecharger

```
1  MetaboliteAdductDecharger
2  -in 2012_02_03_PStd_10_1.featureXML
3  -out_fm MAD_2012_02_03_PStd_10_1.featureXML
4  -algorithm:MetaboliteFeatureDeconvolution:charge_min 1
5  -algorithm:MetaboliteFeatureDeconvolution:charge_max 1
6  -algorithm:MetaboliteFeatureDeconvolution:potential_adducts {H:+:0.7,Na:+:0.3}
```

Further accurate mass search was performed using the extended *AccurateMassSearch* (Listing A.3). Here, the HMDB (4.0) database was used for identification. The database is available in the previously mentioned OpenMS example data. The search space of the positive adducts were reduced to [M+H]+ and [M+Na]+ to fit the adduct grouping step. We used the *id_format* ID to specify the usage of the new ID data structure. In addition, the *mztabm* flag was set to allow the export of the .mztab file in MzTab-M format.

**Listing A.3:** Command: AccurateMassSearch

```
1  AccurateMassSearch
2  -in MAD_2012_02_03_PStd_10_1.featureXML
3  -out MAD_2012_02_03_PStd_10_1.mztab
4  -out_annotation MAD_2012_02_03_PStd_10_1.oms
5  -db:mapping HMDBMappingFile.tsv
6  -db:struct HMDB2StructMapping.tsv
7  -positive_adducts PositiveAdducts.tsv
8  -negative_adducts NegativeAdducts.tsv
9  -algorithm:id_format ID
10 -mztabm
```

To ensure the validity of the generated *MzTab-M* file we used the *Mztab-M* validator jmzTab-m (`https://github.com/lifs-tools/jmzTab-m`, version 1.0.6) to verify the integrity of our file (Listing A.4).

**Listing A.4:** Command: jmzTab-m

```
1  java
2  -jar jmztabm-cli-1.0.6.jar
3  -c MAD_2012_02_03_PStd_10_1.mztab
```

## A.3 Applied Metabolomics: Food Fingerprinting

### A.3.1 Reagents and Chemicals

Ultrapure water was obtained by purifying demineralized water in a Direct-Q 3 UV-R system (Merck Millipore, Darmstadt, Germany). LC-MS grade acetonitrile, methanol as well as ammonium acetate (for analysis EMSURE ACS) were purchased from Merck (Darmstadt, Germany), HPLC grade chloroform was supplied by Carl Roth (Karlsruhe, Germany), LC-MS grade isopropanol and formic acid were provided by Honeywell (Seelze, Germany), ammonium formate solution (10 M in water) was supplied by Sigma-Aldrich (Steinheim, Germany). The reference standards DL-tocopherol palmitate were purchased from Gerbu Biotechnik (Heidelberg, Germany), N-palmitoyl-D-erythro-shinganine (C16 dihydroceramide (d18:0/16:0)) from Avanti Polar Lipids (Alabaster, USA), stearoyl serotonin from Cayman Chemical (Ann Arbor, USA), triolein (TG(18:1(9Z)/18:1(9Z)/18:1(9Z)) from Fluka (Munich, Germany), $\alpha$-tocopherol from Roth (Karlsruhe, Germany), and docosanoic acid tryptamide from Sigma-Aldrich (Munich, Germany).

### A.3.2 Sample Preparation

All cocoa samples were handled identically during the preparations and analytical steps. The sample material was stored at -80 °C prior to sample preparation. Preparation included roasting, separation of the nibs from the shell and the germ, homogenization, and extraction. Extensive analyses have shown that a sample size of 100 cocoa beans (about 120 g) is representative for the respective cocoa sample. Thus, at least 100 cocoa beans were first thawed at room temperature for one hour and roasted afterwards on laboratory scale. The beans were evenly distributed on a grid covered with aluminum foil and roasted at 145 °C in a drying oven (Memmert, Schwabach, Germany) for 30 min. After the beans reached room temperature, the cocoa nibs were separated from the contaminants (shell, germ). Separation was performed manually with the aid of a scalpel. Subsequently, the cocoa nibs and the cocoa shell (shell, germ) were homogenized separately. The homogenization was carried out adding dry ice by means of a Grindomix GM 300 knife mill equipped with a stainless-steel grinding container and a full metal knife (Retsch, Haan, Germany). The homogenate was freeze-dried for 48 hours and stored at -80 °C prior to extraction and analysis. For the extraction, 50 mg of cocoa nibs or cocoa shell lyophilisate were added to the extraction buffer. For the polar extraction, the lyophilisate was subsequently admixed with 900 µL of cooled (4 °C) methanol, two steel balls and dispersed at 3.1 m/s for 2 min in a Bead Ruptor 24 equipped with a 1.5 mL microtube carriage kit (Biolabproducts, Bebensee, Germany). The dispersion was incubated at room temperature for protein precipitation for 5 min. Then, 100 µL of cooled (4 °C) water was added and the sample was treated at 3.1 m/s for cell disruption for 4 min. Both extraction components methanol

and water contained an additive of 5 mM ammonium acetate which was evaluated as the best extraction buffer. After cell disruption, the extraction solution was centrifuged at 14800 rpm at 4 °C for 5 min. The supernatant was membrane filtered using a Rotilabo PTFE syringe filter, 0.45 µm pore diameter (Carl Roth, Karlsruhe, Germany), transferred to a vial, and sealed with a crimp cap. Unless the analysis was performed immediately after extraction, the samples were stored at -20 °C prior to measurement. For the nonpolar extraction, 1 mL of a cooled extraction solution (2-propanol/chloroform (4:1, v/v) + 20 mM ammonium acetate) and two steel balls were added to the weighed lyophilisate. Extraction of the nonpolar metabolites was carried out by a ball mill at 3.1 m/s for 5 min. The following steps of the extraction were analogous to the polar extraction.

### A.3.3 HPLC-ESI-QTOF-MS Data Acquisition

The polar and nonpolar sample extracts were analyzed by a HPLC-ESI-QTOF-MS system using different methods for the two extraction types: the analysis was carried out in the positive and negative ion mode for the polar extracts and in positive mode for the nonpolar extracts. In the experiment, the negative ion mode for the nonpolar samples was omitted, due to sparse-signals detection, which promised no relevance for the cocoa shell detection. In order to counteract environmental and device-related influences, the samples were analyzed in a randomized sequence. Furthermore, in nonpolar analyses, a quality control (QC) sample and a blank (respective extraction solvent) were analyzed every ten samples. Due to the sensibility of the stationary phase, the QC and the blank were measured every five samples during the polar analysis. 20 cocoa nibs samples were worked up as described and merged to form one QC sample. The liquid chromatographic separation of the nonpolar metabolites was performed with a 150 mm x 2.1 mm i.d., 2.6 µm, Accucore RP-MS HPLC column, with a 10 mm x 2.1 mm i.d. guard column of the same material (Thermo Fisher Scientific, Braunschweig, Germany) and a Dionex UltiMate 3000 UPLC System (Thermo Fisher Scientific). The column temperature was set to 40 °C and the flow rate to 350 µL/min. The mobile phase was composed of A (water) and B (isopropanol/acetonitrile) (3:2 v / v). 10 mM ammonium formate buffer (pH 3.5) was added to both eluents. The gradient elution was started at 65% B and kept constant for 2 min., linearly increased to 85% B in 2 min. and afterwards to 100% B in 4 min. 100% B was kept constant for 13 min and was moved back to 6% B in 0.1 min. followed by 3.9 min. of re-equilibration. 5 µL of every sample was injected. For the separation of the polar metabolites a 150 mm x 2.1 mm i.d. 2.2 µm, Cogent Diamond Hydride HPLC column, with a 10 mm x 2.0 mm i.d. guard column of the same material (MicroSolv Technology, Leland, USA) was used with the same HPLC-system. The column temperature was set to 50 °C and the flow rate to 600 µL/min. The mobile phase is composed of the mobile phase A (water) and B (ACN). To both eluents 0.1% of acetic acid was added. The gradient elution was started at 100 % B and

kept constant for 2 min., linearly decreased to 80% B in 2 min., kept constant for 4.5 min., and linearly decreased to 0% B in 2.5 min. 0% B was kept constant for 3.8 min. and was moved back to 100% B in 0.2 min. followed by 4 min. of re-equilibration. 5 µL of every sample was injected. For mass spectrometric analysis, an impact ESI-QTOF (Bruker Daltronics, Bremen, Germany) was used. The analyses were performed in the mass range of 60-1200 Da. The following settings were applied on the system for the nonpolar analysis: end plate offset = 500 V; capillary = 3500 V; nebulizer gas = 4 bar; drying gas = 9 L/min; drying temperature = 200 °C; transition time = 80 µs; pre-pulse storage = 7 µs. Polar analyses were performed on the following device parameters: end plate offset = ± 600 V; capillary = 4000 V; nebulizer gas = 5 bar; drying gas = 9 L/min; drying temperature = 250 °C; transition time = 60 µs; pre-pulse storage = 5 µs. The calibration of the device was ensured by a mixture of formic acid/1 M NaOH in water/isopropanol (0.1:1:100, v/v/v), which was introduced via a syringe pump and a valve switch with a flow rate of 0.1 µL/min directly into the ion source before the re-equilibration step.

### A.3.4 Metabolite Identification and Validation

The identification and determination of the chemical formula were based on the exact mass, isotope ratios, and the analysis of fragment ions. Supplementary, MetFrag[41] was used to create structural suggestions, which were compared to databases such as PubChem[33] and ChemSpider[40]. To prove the hypothetic structural formula, standard commercial substances were analyzed analogously and the retention times, exact mass, isotope ratio, and fragment spectra were compared. However, it was not possible to obtain a standard substance for all compounds. It was ensured to acquire at least one representative substance for each class to compare the characteristic fragmentations with the detected compounds. Especially for fatty-acid tryptamides, fatty acid serotonides, and the tocopherol derivatives, where the compounds differ only in the chain length of the fatty acids, only one or two standard substances were purchased vicariously for all compounds. Nevertheless, a reliable identification can be assumed, since the fragment spectra differs only in the signal for the fatty acid substituent while the other fragment ions are matching. For the identity confirmation of the fatty acid trypamides the docosanoic acid tryptamide was utilized. For the fatty-acid serotonins the stearoyl serotonin and the docosahexaenoyl serotonin were utilized. For the tocopherol derivatives the alpha-tocopherol and the alpha-tocopherol palmitate were utilized. For the ceramide derivatives the N-palmitoyl-D-erthro-sphinganine and the N-(2'-(S)-hydroxypalmitoyl)-D-erythro-sphingosine were utilized and for the triacylglycerols the glyceryl trioleate was utilized.

### A.3.5 Biomarker Identification

Biomarker identification results of method 1 (A.5) and method 2 (A.6).

**Table A.5: Method 1: Results of the Evaluation criteria of the Potential Key Metabolites**

| m/z | Rt [min] | coefficient of variation v [%] | signal areas disparity fermented vs. non-fermented cocoa shell samples [%] | S/N cocoa nibs | S/N cocoa shell/germ |
|---|---|---|---|---|---|
| 144.081 | 9.25 | 15 | 85 | 15 | 90 |
| 161.107 | 9.23 | 12 | 88 | 22 | 105 |
| 161.107 | 9.80 | 14 | 85 | 17 | 142 |
| 197.081 | 10.65 | 59 | 1437 | 4 | 332 |
| 295.227 | 11.95 | 24 | 242 | 11 | 749 |
| 295.228 | 12.02 | 25 | 138 | 7 | 295 |
| 299.239 | 10.62 | 44 | 734 | 42 | 209 |
| 353.307 | 11.37 | 83 | 356 | 5 | 26 |
| 383.368 | 13.50 | 18 | 105 | 19 | 333 |
| 384.347 | 13.50 | 46 | 35 | 18 | 110 |
| 384.347 | 14.08 | 17 | 105 | 7 | 111 |
| 384.347 | 12.28 | 21 | 309 | 5 | 68 |
| 386.400 | 14.77 | 61 | 96 | 8 | 51 |
| 396.368 | 13.45 | 18 | 90 | 14 | 146 |
| 397.384 | 14.45 | 31 | 301 | 6 | 438 |
| 400.416 | 13.60 | 33 | 275 | 7 | 57 |
| 407.319 | 7.32 | 30 | 143 | 10 | 192 |
| 413.378 | 9.93 | 61 | 1202 | 398 | 7206 |
| 427.362 | 8.28 | 85 | 839 | 51 | 1125 |
| 429.332 | 8.88 | 26 | 144 | 42 | 1696 |
| 430.378 | 8.88 | 26 | 147 | 20 | 603 |
| 431.388 | 8.25 | 39 | 150 | 7 | 70 |
| 452.483 | 9.93 | 53 | 403 | 21 | 54 |
| 483.431 | 9.40 | 8 | 93 | 769 | 4356 |
| 469.419 | 8.93 | 22 | 67 | 1785 | 23689 |
| 485.414 | 8.25 | 26 | 68 | 6 | 62 |
| 497.416 | 9.50 | 20 | 79 | 315 | 10198 |
| 499.427 | 8.58 | 27 | 124 | 198 | 4346 |
| 525.481 | 10.08 | 33 | 84 | 192 | 4548 |
| 527.462 | 7.65 | 44 | 115 | 129 | 2369 |
| 529.400 | 7.63 | 36 | 124 | 13 | 147 |
| 529.410 | 9.18 | 20 | 154 | 32 | 305 |
| 536.371 | 9.53 | 21 | 94 | 17 | 121 |
| 540.535 | 10.35 | 38 | 110 | 191 | 4610 |
| 540.535 | 10.63 | 59 | 3294 | 16 | 485 |
| 541.437 | 9.50 | 24 | 124 | 20 | 307 |
| 541.437* | 7.37 | 26 | 70 | 9 | 153 |
| 543.452 | 6.82 | 36 | 164 | 25 | 301 |
| 555.369 | 9.75 | 25 | 175 | 187 | 1088 |
| 556.530 | 10.22 | 53 | 2126 | 37 | 1920 |
| 566.550 | 7.65 | 47 | 60 | 7 | 67 |
| 572.525 | 9.77 | 78 | 6149 | 22 | 449 |
| 597.520 | 9.33 | 84 | 4367 | 37 | 126 |
| 620.577 | 11.37 | 139 | 449 | 19 | 375 |
| 647.572 | 11.38 | 40 | 582 | 113 | 634 |
| 654.603 | 12.60 | 25 | 92 | 6 | 251 |
| 656.595 | 12.50 | 41 | 73 | 1 | 134 |
| 658.613* | 12.75 | 20 | 101 | 12 | 925 |
| 668.598 | 12.25 | 37 | 277 | 673 | 427 |
| 680.637 | 13.50 | 20 | 116 | 15 | 273 |
| 684.631 | 12.70 | 26 | 96 | 1 | 189 |
| 686.648 | 13.13 | 19 | 93 | 30 | 3243 |
| 692.637 | 13.17 | 22 | 127 | 34 | 349 |
| 692.637 | 13.48 | 42 | 84 | 9 | 292 |
| 694.652 | 13.68 | 21 | 109 | 53 | 879 |
| 696.650 | 13.70 | 20 | 113 | 5 | 99 |
| 696.654 | 11.72 | 18 | 232 | 95 | 1433 |
| 698.683 | 15.13 | 52 | 338 | 3 | 15 |
| 708.726 | 13.13 | 18 | 99 | 12 | 113 |
| 713.661 | 13.13 | 39 | 143 | 8 | 108 |
| 714.678 | 13.61 | 22 | 92 | 3 | 170 |
| 720.544 | 8.41 | 83 | 9843 | 5 | 304 |
| 736.756 | 14.05 | 32 | 93 | 4 | 31 |
| 744.592 | 11.30 | 54 | 168 | 14 | 208 |

| | | | | | |
|---|---|---|---|---|---|
| 756.539 | 12.17 | 76 | 3704 | 3 | 111 |
| 764.483 | 8.32 | 74 | 2005 | 4 | 57 |
| 785.703 | 12.31 | 110 | 3997 | 39 | 579 |
| 800.752 | 13.48 | 23 | 122 | 5 | 13 |
| 802.768 | 14.00 | 34 | 111 | 5 | 29 |
| 828.650 | 10.8 | 34 | 225 | 58 | 182 |
| 842.729 | 12.22 | 38 | 396 | 53 | 511 |
| 848.772 | 9.90 | 60 | 186 | 0 | 48 |
| 858.704 | 11.30 | 26 | 190 | 138 | 295 |
| 910.757 | 11.45 | 37 | 46 | 351 | 1030 |
| 940.835 | 13.48 | 33 | 125 | 30 | 160 |
| 942.851 | 13.93 | 41 | 105 | 13 | 548 |
| 944.867 | 14.50 | 36 | 103 | 4 | 154 |
| 946.881 | 15.10 | 37 | 137 | 12 | 18 |
| 956.865 | 14.00 | 167 | 34 | 53 | 5130 |
| 965.856 | 9.25 | 26 | 69 | 46 | 6000 |
| 968.868 | 13.90 | 39 | 102 | 4 | 73 |
| 970.845 | 12.68 | 42 | 53 | 34 | 110 |
| 970.883 | 14.45 | 47 | 98 | 7 | 238 |
| 972.898 | 15.08 | 70 | 79 | 3 | 155 |
| 974.913 | 15.10 | 51 | 140 | 6 | 32 |
| 976.699 | 10.43 | 57 | 92 | 1 | 22 |
| 984.897 | 14.72 | 34 | 118 | 40 | 1024 |
| 990.944 | 14.05 | 97 | 62 | 16 | 28 |
| 991.802 | 9.48 | 38 | 85 | 1 | 23 |
| 998.914 | 15.05 | 45 | 104 | 2 | 87 |
| 1012.929 | 15.38 | 37 | 191 | 31 | 130 |
| 1021.918 | 9.80 | 21 | 83 | 465 | 16247 |
| 1027.825 | 15.77 | 48 | 127 | 3 | 10 |
| 1040.960 | 16.16 | 84 | 74 | 6 | 14 |

**Table A.6: Method 2: Potential Shell Key Metabolites Identified Using *All Relevant Feature Selection***

| **m/z** | **retention time [s]** | **Coefficient of variation (shell)** | **# of missing values (158 shells)** | **Intensity** | **Coefficient of variation (roast)** | **# of missing values (21 roasts)** |
|---|---|---|---|---|---|---|
| 181.071 | 862.44 | 24 | 48 | 95,704 | 22 | 13 |
| 299.237 | 643.21 | 41 | 13 | 77,638 | 20 | 5 |
| 308.294 | 658.62 | 40 | 11 | 93,020 | 19 | 6 |
| 395.366 | 804.52 | 27 | 43 | 220,171 | 10 | 3 |
| 407.316 | 443.35 | 37 | 34 | 185,155 | 60 | 15 |
| 411.398 | 637.16 | 63 | 12 | 85,861 | 15 | 7 |
| 469.364 | 537.13 | 26 | 42 | 215,307 | 59 | 8 |
| 469.416 | 542.79 | 28 | 9 | 439,732 | 52 | 5 |
| 483.432 | 560.25 | 18 | 9 | 38,711,156 | 30 | 5 |
| 483.492 | 560.20 | 30 | 5 | 83,436 | 23 | 3 |
| 484.471 | 575.80 | 74 | 12 | 82,600 | 36 | 7 |
| 485.337 | 537.87 | 21 | 17 | 369,587 | 29 | 9 |
| 497.447 | 578.89 | 25 | 11 | 8,792,394 | 52 | 7 |
| 506.485 | 673.96 | 25 | 8 | 85,210 | 19 | 4 |
| 511.528 | 595.02 | 31 | 6 | 93,596 | 28 | 3 |
| 511.606 | 595.41 | 23 | 17 | 92,050 | 28 | 7 |
| 511.628 | 595.49 | 24 | 6 | 94,414 | 27 | 4 |
| 512.202 | 595.27 | 24 | 13 | 340,293 | 34 | 8 |
| 513.442 | 540.94 | 29 | 12 | 341,495 | 39 | 6 |
| 519.428 | 578.45 | 22 | 8 | 103,928 | 20 | 5 |
| 527.457 | 587.01 | 30 | 13 | 170,183 | 41 | 4 |
| 531.407 | 563.20 | 203 | 10 | 84,395 | 24 | 6 |
| 539.494 | 627.75 | 42 | 8 | 10,438,855 | 69 | 4 |
| 539.502 | 655.02 | 89 | 11 | 93,063 | 41 | 3 |
| 540.534 | 643.10 | 65 | 9 | 421,435 | 65 | 3 |
| 541.473 | 576.88 | 31 | 8 | 281,886 | 62 | 4 |
| 561.477 | 625.41 | 36 | 31 | 111,677 | 41 | 9 |
| 561.487 | 610.21 | 62 | 19 | 155,934 | 44 | 9 |
| 572.525 | 592.13 | 81 | 51 | 281,900 | 66 | 9 |
| 577.519 | 669.24 | 72 | 15 | 1,554,206 | 80 | 5 |
| 578.518 | 578.01 | 32 | 42 | 81,164 | 40 | 8 |
| 583.468 | 611.99 | 61 | 47 | 88,080 | 55 | 14 |
| 594.486 | 616.40 | 36 | 15 | 80,135 | 40 | 7 |
| 595.505 | 563.24 | 89 | 53 | 1,678,039 | 117 | 11 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 595.530 | 669.49 | 65 | 17 | 267,029 | 39 | 9 |
| 599.426 | 514.28 | 18 | 44 | 140,010 | 45 | 14 |
| 601.491 | 602.79 | 59 | 14 | 483,754 | 55 | 10 |
| 601.520 | 647.71 | 73 | 23 | 561,865 | 106 | 4 |
| 605.550 | 690.57 | 153 | 9 | 584,290 | 35 | 5 |
| 607.565 | 713.94 | 142 | 44 | 101,570 | 82 | 12 |
| 608.561 | 694.49 | 91 | 30 | 205,677 | 72 | 6 |
| 617.466 | 599.22 | 65 | 19 | 406,546 | 98 | 3 |
| 617.486 | 564.15 | 61 | 32 | 249,486 | 61 | 8 |
| 620.572 | 687.51 | 38 | 23 | 156,221 | 38 | 1 |
| 623.561 | 692.37 | 138 | 10 | 125,251 | 44 | 6 |
| 628.587 | 663.82 | 44 | 16 | 239,820 | 41 | 11 |
| 640.587 | 694.88 | 168 | 8 | 704,273 | 67 | 4 |
| 647.457 | 690.05 | 35 | 8 | 393,652 | 30 | 4 |
| 647.571 | 654.25 | 84 | 23 | 298,735 | 123 | 9 |
| 653.587 | 766.24 | 27 | 14 | 434,847 | 37 | 9 |
| 656.583 | 703.17 | 98 | 44 | 150,157 | 67 | 17 |
| 656.620 | 680.27 | 146 | 10 | 167,147 | 56 | 8 |
| 658.617 | 774.11 | 24 | 15 | 1,038,459 | 34 | 11 |
| 661.517 | 694.98 | 63 | 12 | 116,301 | 21 | 8 |
| 663.566 | 618.36 | 81 | 33 | 1,237,613 | 84 | 6 |
| 668.597 | 741.81 | 41 | 36 | 122,836 | 32 | 7 |
| 668.618 | 686.86 | 43 | 43 | 494,644 | 103 | 7 |
| 669.618 | 798.70 | 18 | 15 | 688,518 | 8 | 10 |
| 670.631 | 693.70 | 190 | 28 | 84,865 | 44 | 14 |
| 670.636 | 700.83 | 50 | 9 | 973,607 | 48 | 5 |
| 672.613 | 670.17 | 32 | 25 | 95,668 | 14 | 15 |
| 672.631 | 791.13 | 37 | 10 | 338,810 | 49 | 6 |
| 676.530 | 644.38 | 66 | 16 | 367,908 | 61 | 8 |
| 676.547 | 638.94 | 74 | 42 | 84,471 | 101 | 10 |
| 679.541 | 772.50 | 32 | 15 | 88,015 | 24 | 10 |
| 680.689 | 767.82 | 53 | 40 | 103,038 | 57 | 8 |
| 682.599 | 702.77 | 239 | 81 | 271,050 | 91 | 18 |
| 684.627 | 772.51 | 27 | 26 | 216,403 | 58 | 9 |
| 684.628 | 743.31 | 30 | 29 | 110,030 | 11 | 10 |
| 684.652 | 704.97 | 106 | 5 | 943,423 | 81 | 3 |
| 692.632 | 804.15 | 24 | 15 | 257,886 | 27 | 2 |
| 696.649 | 707.04 | 40 | 33 | 1,081,002 | 126 | 9 |
| 700.647 | 692.97 | 46 | 15 | 945,512 | 121 | 8 |
| 701.523 | 624.96 | 58 | 37 | 248,675 | 68 | 6 |
| 707.573 | 799.73 | 22 | 15 | 223,660 | 13 | 10 |
| 708.610 | 700.61 | 44 | 40 | 82,443 | 41 | 14 |
| 712.624 | 697.82 | 61 | 56 | 90,999 | 37 | 11 |
| 712.640 | 694.31 | 23 | 20 | 161,080 | 37 | 8 |
| 712.656 | 753.72 | 70 | 50 | 216,749 | 92 | 8 |
| 712.675 | 719.65 | 52 | 46 | 150,552 | 68 | 9 |
| 712.686 | 731.75 | 87 | 7 | 2,378,312 | 88 | 4 |
| 726.654 | 701.25 | 50 | 58 | 371,779 | 57 | 7 |
| 728.674 | 713.85 | 37 | 16 | 377,730 | 76 | 10 |
| 730.538 | 579.66 | 68 | 43 | 96,152 | 70 | 17 |
| 758.516 | 496.47 | 63 | 18 | 80,850 | 55 | 10 |
| 882.754 | 750.83 | 41 | 33 | 171,900 | 65 | 6 |
| 889.664 | 740.73 | 34 | 8 | 84,700 | 48 | 5 |
| 890.719 | 704.92 | 42 | 35 | 86,085 | 35 | 3 |
| 893.717 | 690.01 | 33 | 6 | 85,438 | 29 | 2 |
| 894.756 | 742.71 | 49 | 7 | 6,612,311 | 104 | 4 |
| 923.830 | 601.22 | 44 | 17 | 134,868 | 42 | 10 |
| 931.674 | 690.47 | 33 | 19 | 92,498 | 20 | 5 |
| 933.690 | 686.28 | 35 | 26 | 133,897 | 37 | 4 |
| 934.694 | 706.50 | 47 | 28 | 108,815 | 39 | 5 |
| 965.851 | 560.62 | 28 | 13 | 4,140,125 | 44 | 8 |
| 993.882 | 578.47 | 35 | 14 | 161,122 | 19 | 9 |
| 997.840 | 520.67 | 42 | 16 | 89,840 | 32 | 9 |
| 1003.806 | 560.40 | 31 | 13 | 107,386 | 24 | 8 |
| 1009.876 | 559.17 | 23 | 16 | 318,158 | 18 | 12 |
| 1059.868 | 595.36 | 21 | 13 | 133,458 | 23 | 8 |

# Appendix B

# Abbreviations

| | |
|---|---|
| *ACN* | Acetonitrile |
| *AGM* | AssayGeneratorMetabo |
| *AMD* | Age-related Macular Degeneration |
| *AMS* | AccurateMassSearch |
| *APM* | Agilent Pesticide Mix |
| *AUC* | Area-Under-the-Curve |
| *Boruta* | All Relevant Feature Selection |
| *CID* | Collision Induced Dissociation |
| *CNV* | Choroidal Neovascularization |
| *CE* | Collision energy |
| *CV* | *MzTab-M*: Controlled Vocabulary, Others: Coefficient of Variation |
| *d-score* | Discriminant-Score |
| *DDA* | Data-Dependent Acqusition |
| *DHA* | Docosahexaenoic Acid |
| *DIA* | Data-Independent Acquistion |
| *EPA* | Eicosatetraenoic Acid |
| *ESI* | Electrospray Ionization |
| *FDR* | False-Discovery Rate |
| *GC-MS* | Gas Chromatography Coupled Mass Spectrometry |
| *GC* | Gas Chromatography |
| *GNPS* | Global Natural Products Society Molecular Networking |
| *GUI* | Graphical User Interface |
| *HMDB* | Human Metabolome Database |
| *HPLC* | High-Performance Liquid Chromatography |
| *HUPO* | Human Proteomoe Organization |
| *IO* | Input/Output |

| | |
|---|---|
| *ID* | Identification |
| *KNIME* | Konstanz Information Miner |
| *LC-MS* | Liquid Chromatography Coupled Mass Spectrometry |
| *LC* | Liquid Chromatography |
| *LDA* | Linear Discriminant Analysis |
| *LOD* | Limit of Detection |
| *m/z* | Mass-to-Charge |
| *MAE* | Mean Absolute Error |
| *MRM* | Multiple Reaction Monitoring |
| *MS* | Mass Spectrometry |
| *MS/MS* | Tandem Mass Spectrometry |
| *MS1* | First stage of mass analysis |
| *MS2* | Second stage of mass analysis (MS/MS) |
| *MSI* | Metabolomics Standard Initiative |
| *MTD* | Metadata Tables |
| *NIST* | National Institute of Standards and Technology |
| *PCA* | Principal Component Analysis |
| *PCV* | Polypoidal Choroidal Neovascularization |
| *PRM* | Parallel Reaction Monitoring |
| *PSI* | Proteomics Standard Initiative |
| *Q* | Quadruople |
| *QC* | Quality Control |
| *qTOF* | Quadruople Time-of-Flight |
| $R^2$ | Coefficient of Determination |
| *RF* | Random Forest |
| *RFE* | Recursive Feature Elimination |
| *RMSE* | Root Mean Square Error |
| *SEH* | Small Molecule Evidence Header |
| *SFH* | Small Molecule Feature Header |
| *SME* | Small Molecule Evidence Table |
| *SMF* | Small Molecule Feature Table |
| *SMH* | Small Molecule Header |
| *SML* | Small Molecule Table |
| *SPLS* | Inverse Sparse Partial Least Squares Regression |
| *SWATH* | Sequential Window Acquisition of all Theoretical Mass Spectra |
| *TOF* | Time-of-Flight |
| *TOPP* | The OpenMS Proteomics Pipeline |
| *UIS* | Unique Ion Signatures |

*XIC* Extracted Ion Chromatogram

# List of Figures

# List of Tables

# Listings

# Appendix C

# Permissions and Contributions

**Chapter 3: Automated, FDR Controlled Targeted Analysis for Data-Independent Acquisition in Metabolomics**[122]

In compliance with the permission and copyright policy, we state that our publication was: "DIAMetAlyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics. Alka O, Shanthamoorthy P, Witting M, Kleigrewe K, Kohlbacher O and Röst H. Nature Communications 13, 1347 (2022)." Licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`). **OA**, HR, OK conceived the project. All authors supplied ideas to the experiment design. MW and KK performed sample preparation and data acquisitions. **OA** developed the method (workflow) and performed the data analysis. **OA** implemented the tools for assay library generation (*AssayGeneratorMetabo*). PS supplied additional experiments (unique ion signatures), discussions and helped in preparing the publication. HR and OK supervised the project. **All** authors discussed the results and contributed to the manuscript. **OA** deposited data to MetaboLights with the identifier MTBLS1108.

**Chapter 4: Reporting Standardization in Metabolomics: *MzTab-M***[6]

In compliance with the permission and copyright policy, we state that our publication was: "mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. Hoffmann N, Rein J, Sachsenberg T, Hartler J, Haug K, Mayer G, Alka O, Dayalan A, Pearce J, Rocca-Serra O, Qi D, Eisenacher M, Perez-Riverol Y, Vizcaíno J, Salek R, Neumann S, and Jones A. Analytical Chemistry 2019 91 (5), 3302-3310. Copyright 2019 American Chemical Society." Further licensed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`).NH, JR, TS,

JH, KH, GM, **OA**, SD, JTM-P, PRS, DQ, ME, YP-R, JAV, RMS, SN, ARJ worked on the *MzTab-M* specification document. **OA** contributed with corrections and adapted examples. NH wrote the paper. **OA** implemented *MzTab-M* data model and data format in OpenMS. **OA** adapted OpenMS tools (*AccurateMassSearch*).

**Chapter 5: Applied Metabolomics: Food Fingerprinting**[15]

Permission to reuse of text, figures, and charts was granted by Elsevier. In compliance with the permission and copyright policy, we state that our publication was: "Food fingerprinting: Mass spectrometric determination of the cocoa shell content (Theobroma cacao L.) in cocoa products by HPLC-QTOF-MS. Nicolas C and Alka O and Segelke T and von Wuthenau K and Kohlbacher O and Fischer M. Food Chemistry, Volume 298, 2019, 125013. Copyright 2019 Elsevier Ltd. All rights reserved." Please direct further permissions related to the material to Elsevier. OK, MF supervised the project. NC, TS, KvW performed the sample acquisition and the experimental analysis. NC performed a part of the post-processing analysis (method 1). **OA** performed the computational MS data analysis, post-processing (method 2) and developed the prediction model. The published manuscript was discussed and edited by all authors.

ARJ: Andrew R Jones; DQ: Da Qi; GM: Gerhard Mayer; HR: Hannes L Röst; JAV: Juan Antonio Vizcaíno; JH: Jürgen Hartler; JR: Joel Rein; JTMP: Jake T M Pearce; KH: Kenneth Haug; KK: Karin Kleigrewe; KvW: Kristian von Wuthenau; ME: Martin Eisenacher; MF: Markus Fischer; MW: Michael Witting; NC: Nicolas Cain; NH: Nils Hoffmann; OA: Oliver Alka; OK: Oliver Kohlbacher; PR-S: Philippe Rocca-Serra; PS: Premy Shanthamoorthy; RMS: Reza M Salek; SD: Saravanan Dayalan; SN: Steffen Neumann; TS: Timo Sachsenberg; TS: Torben Segelke; YP-R: Yasset Perez-Riverol

# Appendix D

# Publications

## Accepted manuscripts

### 2023

Kontou, E. et al. UmetaFlow: An untargeted metabolomics workflow for high-throughput data processing and analysis. (accepted). (2023)

### 2022

Alka, O. et al. DIAMetAlyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics. Nat Commun. (2022)

### 2021

Morgenstern, M. et al. Quantitative high-confidence human mitochondrial proteome and its dynamics in cellular context. Cell Metab. (2021)

Bichmann, L. et al. DIAproteomics: A multifunctional data analysis pipeline for data-independent acquisition proteomics and peptidomics. J. Proteome Res. 20, 3758–3766 (2021)

### 2020

Alka, O. et al. CHAPTER 6: OpenMS and KNIME for Mass Spectrometry Data Processing. Processing Metabolomics and Proteomics Data with Open Software: A Practical Guide. R. Soc. Chem., 201–231 (2020).

Scheidt, T. et al. Phosphoproteomics of short-term hedgehog signaling in human medulloblastoma cells. Cell Commun. Signal. 18, 99 (2020)

Rurik, M., Alka, O., Aicheler, F. & Kohlbacher, O. Metabolomics Data Processing Using OpenMS. Methods Mol. Biol. 2104, 49–60 (2020)

Kutuzova, S. et al. SmartPeak automates targeted and quantitative metabolomics data processing. Anal. Chem. 92, 15968–15974 (2020)

Nothias, L.-F. et al. Feature-based molecular networking in the GNPS analysis environment. Nat. Methods 17, 905–908 (2020)

## 2019

Cain, N. et al. Food fingerprinting: Mass spectrometric determination of the cocoa shell content (Theobroma cacao L.) in cocoa products by HPLC-QTOF-MS. Food Chem. 298, 125013 (2019)

Licha, D. et al. Untargeted Metabolomics Reveals Molecular Effects of Ketogenic Diet on Healthy and Tumor Xenograft Mouse Models. Int. J. Mol. Sci. 20, (2019)

Hoffmann, N. et al. MzTab-M: A data standard for sharing quantitative results in mass spectrometry metabolomics. Anal. Chem. 91, 3302–3310 (2019)

## 2018

Peters, K. et al. Current Challenges in Plant Eco-Metabolomics. Int. J. Mol. Sci. 19, (2018)

## 2017

Pfeuffer, J. et al. OpenMS – A platform for reproducible analysis of mass spectrometry data. J. Biotechnol. 261, 142–148 (2017)