Novel analytical and data analytical approaches for *Caenorhabditis elegans* metabolomics and lipidomics

Habilitationsschrift

vorgelegt von

Dr. rer. nat. Michael Witting

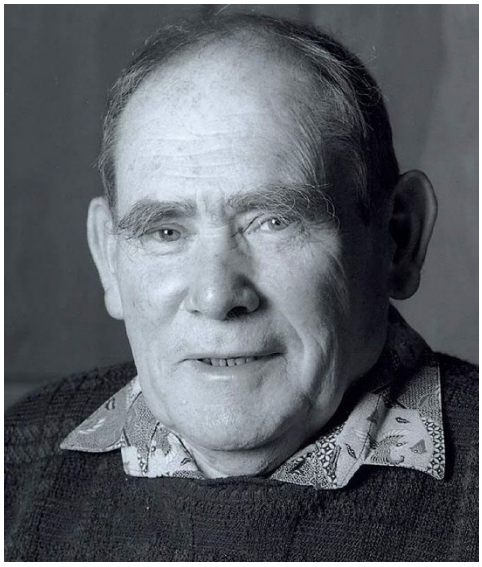
Lehrstuhl für Analytische Lebensmittelchemie

TUM School of Life Sciences

Freising

Technische Universität München

2023



© Getty Images

As was predicted at the beginning of the Human Genome Project, getting the sequence will be the easy part as only technical issues are involved. The hard part will be finding out what it means, because this poses intellectual problems of how to understand the participation of the genes in the functions of living cells.

Sydney Brenner (1995)

To Fanny and Jack!

Acknowledgement

First and foremost, I would like to thank my wife. My life would be nothing without you! Thanks for all the support over all the last years. Thanks to our little boy Jack for coming unexpectedly into our life and being a source of joy and happiness ever since then.

I would like to thank my family for all the support throughout the years, during school, my studies, PhD and finally this chapter. Thank you very much Mama, Papa, Toni and Hanna as well as Hermann, Evi and Dani.

Thanks to all the colleagues, coworkers, collaboration partners and friends over the last years! There are too many to list them all.

My special thank goes to Philippe for all the support during my PhD, PostDoc and habilitation. Thank you very much for giving me the freedom to pursue my own academic path.

Thanks to Prof. Michael Rychlik and Prof. Martin Klingenspor for serving on my habilitation committee.

List of Abbreviations

ACN	Acetonitrile
APCI	Atmospheric Pressure Chemical Ionization
BCAA	Branched-chain Amino Acids
BCFA	Branched-chain Fatty Acids
C. elegans	*Caenorhabditis elegans*
CCS	Collisional Cross Section
CE	Cholesteryl Ester
Cer	Ceramide
ChEBI	Chemical Entities of Biological Interest
DDA	Data-Dependent Acquisition
DG	Diacyl Glycerol
DIA	Data-Independent Acquistion
DNA	Desoxyribonucleic acid
DTIMS	Drift Tube Ion Mobility Spectrometry
ECMDB	Escherichia coli Metabolome Database
EI	Electron Ionization
ESI	Electrospray Ionization
FT-ICR-MS	Fourier Transform Ion Cyclotron Resonance Mass Spectrometry
GalCer	Galactosylceramide
GC	Gas Chromatography
GluCer	Glucosylceramide
GNPS	Global Natural Products Social Molecular Networking
GPNAE	Glycerophospho-N-Acylethanolamide
HexCer	Hexosylceramide
HILIC	Hydrophilic Interaction Liquid Chromatography
HMDB	Human Metabolome Database

iPrOH	iso-Propanol
L1	L1 larvae
L2	L2 larvae
L3	L3 larvae
L4	L4 larvae
LC	Liquid Chromatography
logP	2-Octanol-Water partition coefficient
LPC	Lysophosphatidylcholine
LPE	Lysophosphatidylethanolamine
LPI	Lysophosphatdiylinositol
m/z	Mass-to-charge ratio
MeOH	Methanol
MG	Monoacylglycerol
MS	Mass Spectrometer / Mass Spectrometry
MTBE	Methyl-tert-Butyl ether
MUFA	Monounsaturated Fatty Acids
NAE	N-Acylethanolamide
NAPE	N-Acylphosphatidylethanolamide
NAPS	N-Alkylpyridinium-3-sulfonate
NMR	Nuclear Magnetic Resonance Spectroscopy
PA	Phosphaditic acid
PC	Phosphatidylcholine
PCA	Principal Component Analysis
PE	Phosphatidylethanolamine
PI	Phosphatidylinositol
PS	Phosphatidylserine
PUFA	Polyunsaturated Fatty Acids
QC	Quality Control

RI	Retention Index
RNA	Ribonucleic acid
RP	Reversed-phase
RSD	Relative standard deviation
RT	Retention time
RTI	Retetion time indexing
SFC	Supercritical Fluid Chromatography
TG	Triacylglycerol
TIMS	Trapped Ion Mobility Spectometry
TLC	Thin Layer Chromatography
TWIMS	Traveling Was Ion Mobility Spectrometry
UHPLC	Ultrahigh performance liquid chromatography
UHR-ToF-MS	Ultra high resolution-Time of Fligh-Mass Spectrometry
YA	Young Adult
YMDB	Yeast Metabolome Database

Summary

The small nematode *Caenorhabditis elegans* (*C. elegans*) is one of the premier model organisms in biomedical research. Metabolomics and Lipidomics have been recently added to the toolbox to explore the worm's biology. Despite years of research in *C. elegans* metabolism, only a small part of the metabolome and lipidome is known, and even fewer metabolites can be annotated in typical metabolomics experiments. The correct annotation and identification remain one of the major bottlenecks in metabolomics, possessing a problem not specific to *C. elegans*. In order to make metabolomics and lipidomics more attractive as a functional read-out for *C. elegans* experiments, several obstacles need to be overcome.

This work summarizes work conducted by supervised bachelor, master, and PhD students and me to develop solutions for different parts of the workflow in *C. elegans* metabolomics and lipidomics. First, the exact number of metabolites and lipids in the nematode remains unknown. A genome-scale metabolomics model called WormJam was developed based on previously published models to form a resource on currently known metabolic pathways in *C. elegans*. This model was further corrected and enlarged with *C. elegans*-specific reactions to form a rich knowledge base, which also includes information on metabolites and lipids present. Second, metabolite and lipid identification approaches have been developed to enable the systematic and reproducible analysis of metabolomics datasets. This included retention time indexing as a new approach to normalize retention information and employing ion mobility and analysis of tandem MS data in the case of lipids. Third, the structural diversity of lipids is often underappreciated and large-scale identification to generate a blueprint for the C. elegans lipidome has been performed, which included the curation of lipids from literature as well as the in-depth analysis of different obtained C. elegans lipidomics datasets. Lastly, new analytical approaches are required to delve deeper in the metabolome and lipidome. A tandemLC setup has been developed to allow the combined analysis of polar and non-polar metabolites from a single injection. Furthermore, a method

for analyzing fatty acids from single worms has been developed, paving the way toward single-worm analysis of metabolomes.

All methods together will allow in the future a more comprehensive analysis of the C. elegans metabolome and lipidome, especially from a lower number of worms. This represents a significant development to integrate metabolomics and lipidomics in large-scale experiments, such as genetic screens.

Content

- This page is left intentionally blank -

1. Introduction

Chapter-related publications:

Quo Vadis Caenorhabditis elegans Metabolomics-A Review of Current Methods and Applications to Explore Metabolism in the Nematode

Salzer L., M. Witting

Metabolites. 2021 Apr 29;11(5):284. doi: 10.3390/metabo11050284

This review summarizes the current state of the art in *C. elegans* metabolomics and lipidomics. It was written by my Ph.D. student Liesa Salzer and myself in collaboration. As part of this review, we also performed a curation of metabolites found in different publications.

1.1. *Caenorhabditis elegans* – A versatile model organism

The nematode *Caenorhabditis elegans* (*C. elegans*) was first described by Emile Maupas as *Rhabditides elegans* and further studied by Victor Nigon on reproduction, meiosis, and development [1-3]. In 1952 it was placed into the subgenus *Caenorhabditis,* and in 1955 this was raised to the genus level. The name is derived from Greek and Latin words and means elegant and rod-like (rhabditis = rod-like, elegans = elegans, caeno = recent). It was introduced in 1969 by Sidney Brenner as a model organism for development and genetics [4]. *C. elegans* is among the most commonly used laboratory animals and one of the easiest to handle.

C. elegans is found in humid, temperate areas virtually globally [5, 6]. It is typically isolated as a stress-resistant dauer stage (see below) from rich soil and rotten fruits [7, 8]. The worm has a boom-and-bust lifestyle and can migrate over short distances, but longer distances are possible through hosts like birds, rodents, or humans [9]. In the wild several bacterial genera, such as *Pseudomonas*, *Stenotrophomonas*, *Ochrabactum,* and *Sphingomonas* are associated with *C. elegans* forming its native microbiome [10]. Sterols required by the worm are potentially supplied by yeast or rotten plant material in the wild.

In the laboratory, *C. elegans* is cultivated on solid agar plates monoxenically using *Escherichia coli*, another model organism, readily available in many laboratories at the time of introduction, as food source. Since *C. elegans* is auxotrophic on sterols, cultivation plates must be supplemented with a suitable sterol source, such as cholesterol. Usually, the uracil auxotrophic *E. coli* strain OP50 is used, which was introduced by Sydney Brenner since it grows in thin lawns allowing better microscopic visualization of *C. elegans* [4].

C. elegans has several nutritional requirements. Typical *E. coli* fed to *C. elegans* consists of 55% protein, 23% nucleic acids, 7-9% lipids, 6% carbohydrates, and 4% vitamins, cofactors and ions (dry weight, numbers from [11]). The exact biomass of *C. elegans* has not yet been characterized but differs from the *E. coli* food source and is closer to mammalian compositions. It comprises roughly 60% protein, 20% lipids, 6.5% nucleic acid, and 6%

carbohydrates (dry weight). Examples of essential nutrients for *C. elegans* are the amino acids arginine, histidine, lysine, tryptophan, methionine, phenylalanine, leucine, isoleucine, valine, and threonine, heme, and sterols, as well as different B vitamins. Detailed nutritional requirements have been reviewed by Zečić et al. [11]. Although genetic approaches and research are well-developed in *C. elegans*, biochemistry lacks a bit behind, primarily due to the lack of axenic cultivation possibilities. In purely axenic media, worms show slow growth and development. However, the addition of particulate matter improved growth [12]. New biochemical and systems biology approaches like transcriptomics, proteomics, metabolomics, and lipidomics are increasingly applied and will help close this gap and readily support nutritional studies.

C. elegans has two sexes, hermaphrodites (⚥) and males (♂); both sexes are diploid and have five autosomal chromosomes. Hermaphrodites have two X chromosomes (XX), while males have only one (XO). Males occur by spontaneous non-disjunction of the X chromosome at a low frequency of 0.1-0.2% or higher frequencies (up to 50%) through mating. The life cycle is fast, roughly three days at 25°C from fertilized egg to adult. At 20°C, embryogenesis takes about 16 hours. The first stages of development take place within the mother and are independent of her due to a practically impermeable eggshell. Eggs are typically retained within the mother until the 24-cell stage. Upon hatching, the first larval stage (L1) is produced, which has an approximate duration of 16 hours. All following larval stages (L2-L4) last about 12 hours (see Figure 1). Under harsh environmental conditions, the worm can enter an alternative development stage called the dauer stage (from German “dauer” for enduring), which can survive up to several months. Once conditions ameliorate, the dauer stages develop through L4 larvae to normal adults without compromise in total life span. *C. elegans* develops from fertilized eggs through different larval stages into adult animals capable of laying eggs again.

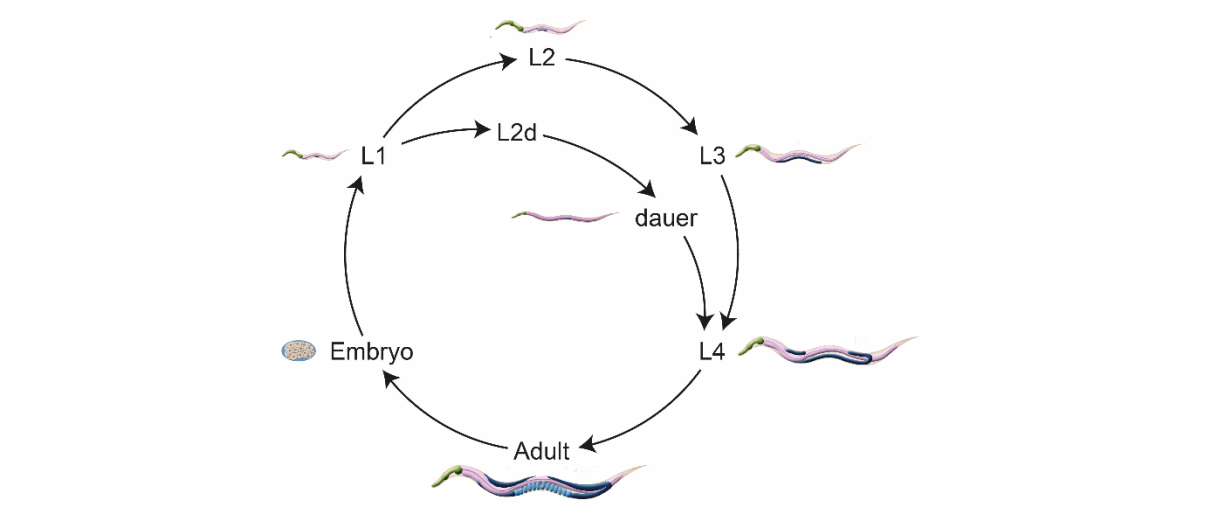


Figure 1: C. elegans develops into reproductive adults through four larval stages. Under harsh environmental conditions, L1 larvae can enter an alternative live cycle and develop into dauer larvae, which can survive several months without feeding. Once conditions improve, they develop into L4 larvae.

The hermaphrodite produces sperm and stores them in spermatacea in the L4 stage. During selfing, hermaphrodites use the sperm produced by themselves. About 200-300 eggs are laid, while when mating with males, ~1000 worm offspring can be made, indicating that sperm produced by the hermaphrodite is the limiting factor. Twelve hours after the L4 to adult molt, hermaphrodites begin to lay eggs for two to three days until all self-produced sperm is used. However, sperm-depleted hermaphrodites can still mate with males. After their reproductive phase, worms live for several weeks before dying of senescence. The development of *C. elegans* is invariant, and the complete cell lineage is known [13-15]. The adult hermaphrodites consist of 959 somatic cells, and the male of 1031 cells contain defined tissues. Of these 959, 302 cells are neurons belonging to two distinct and independent nervous systems. The worm was the first unicellular organism with a completely sequenced genome [16]. In 2019 the complete connectome of both sexes was published, making *C. elegans* the first organism with a wholly known neuronal wiring diagram [17]. The transparent body makes it easy to study cellular and subcellular details, e.g., visualized by Nomarski microscopy.

Several features make the nematode a widely used model organism employed in biomedical research, e.g., aging, neurobiology, development, host-pathogen and host-microbe interactions, and metabolism [18-21]. The small size, large brood size, ease of cultivation,

low maintenance, long-term cryopreservation, invariant cell number and development, known cell lineage, and several available genetic tools make *C. elegans* an ideal model organism for basic and applied research. Selfing and mating hold unique opportunities for genetic studies. Any non-lethal mutation can be maintained through self-propagation, and populations are essentially isogenic. Furthermore, since selfing follows the standard Mendelian segregation rules, parental heterozygous strains of a recessive trait will produce the standard 1:2:1 pattern, meaning that 25% of the progeny will be homozygotes of the mutant allele and display the recessive trait. Crossing of different genetic backgrounds is easily achievable through mating. Since male sperm is used before hermaphrodite sperm, defined crossings can be obtained. Lastly, *C. elegans* suffers not from inbreeding depression [22].

Several genetic tools exist in *C. elegans,* and the possibility of maintaining a genetically constant population in a relatively simple cultivation system allowed large-scale genetic investigations [23-27]. The transparent body enables the localization of different dyes or fluorescent proteins and protein fusions in the body. RNAi enables the silencing of genes at specific time points and allows to study also the function of genes for which the total loss would be lethal. CRISPR/Cas-9 is also used in *C. elegans* for genetic modification. From the *C. elegans* genome, about 38% have a predicted human ortholog, while 60-80% of human genes have an ortholog in the worm [28, 29]. About 40% of human disease genes have an ortholog in *C. elegans* [30]. However, also several drawbacks exist. For example, many genes of the Hedgehog signaling pathway are missing [31]. Furthermore, no cell culture lines of *C. elegans* cells exist, and the axenic culture of the nematode is only poorly developed. Currently, WormBase contains 48433 genes, of which 20184 are protein-coding and 26720 are non-coding RNA and pseudogenes [32].

Table 1: Comparison of different model organisms (adapted from [33]). Distinct advantages (+) and disadvantages (-) are listed and compared. The genome size and the number of protein-coding genes are listed for relative comparison.

Organism	Advantages / Disadvantage	Genome Size (Mb)	Number of protein-coding genes
Mus musculus (mouse)	(+) Genome is available (+) Strong genetic, physiological overall with humans (-) Ethical concerns (-) Expensive (-) Long generation time (2-3 months) (-) Not amenable to HT screens	2689.66	20,000
Drosophila melanogaster (fruity fly)	(+) Inexpensive/easy to grow (+) Genome is available (+) Straightforward genetic tools exist (+) short generation time (~ 10 days) (+)/(-) 50-80% of fly genes homologous to human genes	137.688	14,000
Caenorhabditis elegans (round worm)	(+) Inexpensive/easy to grow (+) genome is available (+) Straightfoward genetic tools exist (+) Short generation time (2-3 days) (+) short lifespan (2-3 weeks) (+) small, exactly 959 somatic cells (+) invariant development (+) transparent (+) has organs / differentatiated tissues (+) mutants can be frozen (+)/(-) 50-80% of worm genes homologous to human genes	102.042	20,000
Danio rerio (zebrafish)	(+) draft genome is available (+) small (+) embryos are transparent (-) long generation time (2-4 months) (-) isogenic strains are not available	1427.29	26,200
Saccharomyces cerevisiae (budding yeast)	(+) simple model of eukaryotic cell (+) genome is available (+) straightforward genetic tools exist (+) simple development (sporulation) (+) some intercellular interaction (mating)	11.8643	6600

Organism	Advantages / Disadvantage	Genome Size (Mb)	Number of protein-coding genes
Escherichia coli	(+) inexpensive/easy to grow (+) genome is available (+) straightforward genetic tools exist (+) best studied organism (+) model for molecular genetics (-) limited differentiation (-) limited intercellular interaction	5.13951	4300

1.2. Systems Biology, Metabolomics and Lipidomics

1.2.1. Definition of Systems biology

Living biological systems represent complex systems. These complex systems are built of many different building blocks which interact to make up life and carry out biological functions. Only the concerted action of all these building blocks can carry out correct functions. Systems biology aims to understand biological systems or organisms as a whole. Its goal is to integrate knowledge of regulatory processes on all levels and be able to predict potential outcomes upon perturbation of a given system. Systems biology combines approaches used in biology with mathematical modeling and bioinformatics to deduce novel knowledge and be able to predict future outcomes of treatment, genetic modification, etc.

In contrast to the reductionist view of molecular biology, systems biology is a holistic approach. In order to fulfill its promises, it is necessary to quantify all biological entities and their interactions in a given system at a given time, which includes DNA, RNA, proteins, and metabolites. In the classical view of biology, information flows from DNA via RNA to proteins, which then execute biological functions. Metabolites are converted into each other by enzymes encoded in the genome, providing energy and building blocks for an organism to survive. Quantitative information can be obtained by using so-called “omics” approaches genomics, transcriptomics, proteomics, and metabolomics. In early 2010 Sydney Brenner stated in his article “Sequences and consequences”: “The new science of Systems Biology

[…] will fail because deducing models of function from the behavior of a complex system is an inverse problem that is impossible to solve.” [34]. Systems biology is often believed to be only possible using high throughput techniques generating a lot of data which is then used in modeling approaches to deduce novel knowledge, which was the main criticism of Sydney Brenner. However, systems biology integrates a lot of molecular biology knowledge created before in the models and relies on the foundation of molecular biology and biochemistry.

Although the genome as the blueprint of life is known, our understanding of how a living organism works is minimal. It requires a detailed understanding of each molecular entity and its action and the concerted actions of all entities. This level is not yet reached in biology. Lazebnik was working on apoptosis and compared the way biological knowledge is generated to the way electrical engineers work [35]. While engineers know the function of each part and how they work together. Once biology reaches this level of understanding, we can predict outcomes of mutations, changes in the environment, etc., and how they relate to health and disease.

One way to reach this level of understanding is to employ the aforementioned “omics”-technologies. The different "omics" aim to characterize DNA and its modification, changes in transcription and RNA, translation and protein levels, and metabolism comprehensively. While DNA represents a relatively static blueprint, metabolism is highly dynamic. Therefore, the study of metabolism is of great interest since, from the attributes defining life, it is one of the most important since it delivers energy and building blocks or acts as signaling molecules. Furthermore, metabolism is typically one of the first to react to an external stimulus.

1.2.2. Definition and History of Metabolomics

Metabolomics focuses on the holistic study of metabolism in a given biological system. Metabolism has been studied for centuries, but only recent technological advances made it possible to study the metabolism more holistically. The term “metabolome” was first time mentioned by Oliver *et al.* in 1998 and describes in accordance with the genome (the entire

set of all genes), all metabolites present in a living system [36]. Different definitions of metabolomics exist. Fiehn defined metabolomics as “[…] comprehensive analysis in which all the metabolites of a biological system are identified and quantified […]” [37]. Nicholson defined the field of metabonomics as “[…] the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification.” [38]. Metabolomics flourished in the last years because of the development of newer and better analytical systems and the development of bioinformatics approaches, which made it possible to cover more metabolites in a single analysis.

The different “omes” are very different based on the underlying chemical structures and physicochemical properties. While DNA, RNA, and proteins are linear polymers of defined building blocks, metabolites do not follow such a large-scale organizational principle (Figure 2). DNA consists of a backbone made from deoxyribose and phosphate groups. Attached to the sugar are four different bases, adenine, cytosine, guanine, and thymine. The DNA is organized in a double helix and nucleotides from the characteristic base pairs, A and T and G and C, which interact based on either 2 or 3 hydrogen bridges. RNA consists of a single strand based on ribose and phosphate groups and contains uracil instead of thymine. Analytical tools for DNA and RNA analysis are based on the same underlying principle: complementarity of strands and hydrogen bonding. Furthermore, DNA and RNA can be amplified from limited starting material. Proteins are based on 20 different proteinogenic amino acids, which form a linear polymer of amide bonds with different residues. The functional diversity of proteins is derived from the different 3D structures that can be formed based on the amino acid sequence. In contrast to DNA and RNA, proteins cannot be amplified.

Metabolites don’t follow such a common structural motif and cover a large range of physicochemical properties ranging from very polar metabolites (e.g. tri methylamine-N-oxide TMAO, logP = -2.57) to non-polar lipids (e.g. tristearine TG(18:0/18:0/18:0), logP = 21.59). Each metabolite has its own unique chemical structure, which leads to a large complexity of

the metabolome. Single reactions converting metabolites into each other can greatly impact the molecular structures and properties. For example, a transamination reaction produces glutamine by transferring an amino group from the amino group of glutamate to produce the amid group in glutamine. Although metabolites show some core structures, their properties are highly different. For example, while the amino acids glutamic acid and glutamine represent zwitterionic structures, 2-oxo-glutarate is an acid that has no iso-electric point. Also, their logP values are very different (see Table 2).

Table 2: Comparison of Glutamic acid, Glutamine, and 2-oxo-glutaric acid, all members of the same reaction sequence. The Tanimoto similarity is a measure of the chemical similarity. Even within this single reaction sequence, metabolites show differences in their chemical properties, exemplifying the problem of large chemical complexity metabolomics has to deal with.

Metabolite	Tanimoto similarity			pI	logP
	Glutamic acid	Glutamine	2-Oxo-glutaric acid		
Glutamic acid	1	0.72	0.46	2.80	-3.24
Glutamine	0.72	1	0.37	5.73	-4.00
2-Oxo-glutaric acid	0.46	0.37	1	---	-0.11

Although this is only a relatively simple example, it demonstrates the large variety a single reaction can introduce in the properties of metabolites. Due to the extensive structural variation of metabolites and the lack of standardization in metabolomics, it is still regarded as “complicated”. No single analytical method is capable of covering the entire metabolome. Additionally, the large polarity range of metabolites can span several orders of magnitude in concentration. Glucose can be found even in millimolar concentrations in plasma for example, while potent signaling metabolites are only present a very low concentrations, e.g. endocannabinoids like anandamide [39, 40].

In order to tackle the complexity of the metabolome and lipidome, different analytical approaches are required. Mass Spectrometry (MS) and Nuclear Magnetic Resonance spectroscopy (NMR) represent the two most commonly employed methods for the analysis of the metabolome/lipidome, which are described below [41].

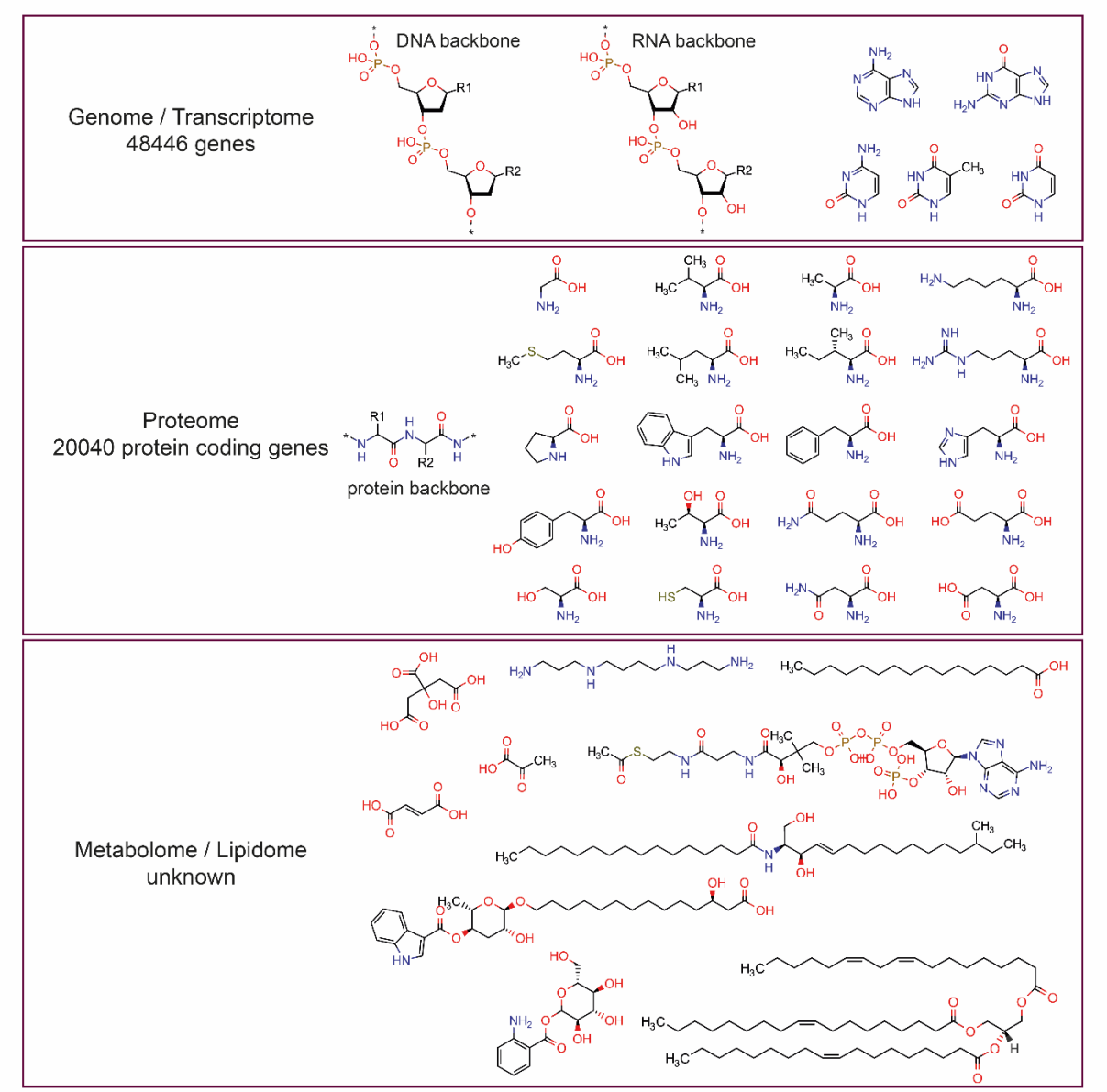


Figure 2: The genome, transcriptome, proteome, metabolome and lipidome are following different chemistries and different in their complexity. While for DNA, RNA and proteins, a common underlying chemical structure is present, for metabolites and lipids this is not the case. DNA, RNA and proteins represent linear polymers with repeating units (referred in the figure as backbone). For analysis of these molecules are restricted chemical space is covered and utilizes common physicochemical principles (e.g. hybridization of complementary strands for DNA or RNA) or the restricted chemical spaces (peptides in case of proteins). Metabolites and lipids in contrast show a high diversity of structures, which cannot be analyzed by a single analytical method.

1.2.3. NMR based metabolomics

NMR spectroscopy is based on the interaction between the magnetic moment of atomic nuclei and an external magnetic field and their perturbation by a weak oscillating electromagnetic field. First, the nuclei are polarized by the external, constant magnetic field, followed by perturbation of this polarization using an oscillating electromagnetic field. Then, the relaxation of the perturbed nuclei is measured due to an induced current in a detection

coil. The free induction decay (FID) is obtained, which is the magnetic resonance response in a unit of time. The NMR spectra are generated via Fourier-transformation (FT), consisting of signal intensities of NMR active compounds depending on their resonance frequency. Only nuclei that have a non-zero nuclear spin do have a magnetic moment and thus are NMR active. The most important nuclei in the analysis of biomolecules are ^{1}H, ^{13}C, ^{15}N, ^{19}F, and ^{31}P. The sensitivity of NMR analysis is influenced by the natural abundance of the nuclei, e.g. ^{13}C is only present at 1.1%. Thus ^{13}C-NMR requires longer times.

In metabolomics, mostly one-dimensional ^{1}H NMR spectroscopy is used since it is fast and very convenient for the universal detection of organic compounds. NMR is quantitative and nondestructive, with minimal sample preparation and interference, resulting thus to the lowest analytical variation compared to other techniques [42]. Its high precision allows the detection of even small changes in metabolite abundances. NMR delivers qualitative (structure) and quantitative information in a single run. Its strengths are particularly evident in substances that are difficult to ionize in MS or require derivatization. However, a significant disadvantage of NMR compared to MS is the lower sensitivity. Therefore only a few tens to hundreds of high abundant compounds are covered by 1D-^{1}H NMR [43]. Besides profiling, NMR is often used in metabolite identification for structure elucidation. It's often combined with Liquid Chromatography (LC) in an online or offline (fractionation) approach to lower the complexity of the individual samples and can be combined with MS [44, 45].

1.2.4. MS-based metabolomics

MS is a premier tool in (bio)medical and chemical research, environmental, forensic, and medical analysis. A mass spectrometer is a device that ionizes substances and determines their mass-to-charge ratio (*m/z*). MS is beside NMR, one of the major analytical technologies used in metabolomics. Ions are separated due to different physical approaches, depending on the type of mass spectrometer being used. In MS, *m/z* ratios of ions are measured together with their corresponding intensities, combining both qualitative (*m/z*) and quantitative (intensity) information. However, metabolites with the same molecular formula

have the same mass and, therefore the same *m/z* value and are called isomers (e.g.$[M+H]^+$ Leucine and Isoleucine, $C_6H_{13}NO_2$, *m/z* 132.1019). Likewise, several ions might have very similar but not identical *m/z* values (isobars, e.g. $[M+Na]^+$ adduct of PC(34:1), *m/z* 782.5670 and $[M+H]^+$ adduct of PC(36:4), *m/z* 782.5694). Their separation depends on the resolution of the employed MS instrument. Metabolites may be further analyzed using tandem MS (sometimes called MS^2 or MS/MS) experiments. Here, the ions of interest are fragmented, mostly using collision-induced dissociation (CID), and fragment ion *m/z* values are analyzed, yielding information on potential substructures. This allows to potentially differentiate isomeric and isobaric structures. In non-targeted metabolomics, data-dependent acquisition (DDA) is used to generate and acquire MS/MS data. In DDA the n most intense ions in an MS^1 scan, meeting specific user-defined thresholds, are selected and fragmented in the following n scans. Coverage of metabolites selected for fragmentation depends on the scan speed and duty cycle of the used MS instrument (typically top 5 to 25 are used). Another approach that is increasingly used in non-targeted metabolomics is data-independent acquisition (DIA), such as Sequential Window Acquisition of all Theoretical fragment ion spectra (SWATH) MS or All ions [46]. These techniques don't rely on a previous selection of precursor *m/z* but fragment either all ions at the same time or scan the *m/z* range in defined windows. The benefit of DIA is that there is no need to repeat analysis when interested in different compounds since every precursor was already fragmented, which allows retrospective data analysis without the need of reacquisition.

Different types of mass analyzers exist, and they can be combined in different ways. The most commonly used MS in metabolomics are triple quadrupole MS (QqQ) for targeted analysis and quantification and quadrupole time of flight MS (qTOF), Orbitrap or Quadrupole-Orbitrap and quadrupole Fourier transform-Ion cyclotron resonance-MS (qFT-ICR) for untargeted analysis. qToF, Orbitrap, and FT-ICR-MS offer very high or even ultra-high resolution. This high resolution enables better separating different *m/z* values and better determining potential sum formulae candidates. Together with an accurate isotope pattern

detection, this enables the calculation of sum formulae, improving searches in databases compared to accurate mass alone [47]. Metabolite identification benefits not only from high mass resolution and accuracy in full MS analysis but also during fragment analysis.

MS is either employed without or with prior chromatographic or electrophoretic separation. The employed ionization source depends on the upfront sample introduction system, e.g. electron ionization (EI) or chemical ionization (CI) for Gas Chromatography (GC), electrospray ionization (ESI), and atmospheric-pressure chemical ionization (APCI) for direct infusion, Liquid Chromatography (LC) or Capillary Electrophoresis (CE). MS hyphenated to a separation method reduces the complexity of mass spectra, providing separation of isomeric and isobaric structures and additional information about physicochemical properties of detected metabolites based on the respective separation characteristics. The use of DI- or FIA-MS allows high-throughput analysis of metabolomes, e.g. of a genome-wide metabolomics screen in *E. coli* [48]. The limitation of this approach is the missing separation of isomeric structures.

Chromatography is based on the differential partitioning of metabolites between a mobile and stationary phase. Based on the employed mobile phase different methods are differentiated. GC uses a gas, typically helium or hydrogen, as mobile and a thin immobilized liquid file as stationary phase. GC was one of the earliest adopted analytical technologies for the analysis of the metabolome, although at this time this term was not existent. Pauling et al. applied GC for the analysis of endogenous substances in human urine and breath [49]. Separation of metabolites in GC is based on the different volatility and portioning of substances. Since the separation takes place in the gas phase, analytes of interest have to be volatile and need to be made volatile by derivatization. This is typically the case for metabolites with a mass lower than 400 Da. Separations in GC show a high efficiency and narrow peaks. Furthermore, a wide range of metabolites can be analyzed on a single stationary phase. Dependent on the employed MS instrumentation, different types of information can be obtained, ranging from nominal masses and molecular formulae to fragmentation patterns. Coupling with MS is

achieved since molecules are already in the gas phase and must only be ionized. EI is the premier choice for ionization in GC-MS. Metabolites are ionized by the impact of a high energetic electron, typically 70eV. Since this energy is much higher than typical chemical bonds, this process leads to not only ionization but also fragmentation of the molecule. Often the parent ion cannot be seen or is of very low intensity. However, the fragmentation process under EI conditions is highly reproducible, and libraries of fragmentation spectra are available and universally useable, independent of the instrument vendor. This improves metabolite identification and is a major advantage of GC-MS. GC-MS involves the additional step of metabolite derivatization, which can increase variability. Furthermore, the derivatization analysis of GC-MS also involves several more steps for data analysis.

If metabolites are not volatile or cannot be made volatile by derivatization LC can be employed for separation. LC utilizes different solvents and stationary phases (typically based on a silica gel support), and different separation methods are available. Normal phase (NP) separations use silica gel and apolar organic solvents like hexane or chloroform. This separation method is not widely applied but has some advantages for lipid analysis. Reversed-phase (RP) separation is the most employed separation method in metabolomics. The name relates to the reversed polarity compared to NP with an apolar stationary phase (e.g. octadecyl modified silica gel, C18) and polar mobile phases like water, MeOH, or ACN, and their mixtures. This method is well suited for analyzing non-polar metabolites like fatty acids, acyl-carnitines, or secondary metabolites. To overcome this, methods such as ion-pair chromatography (IP) have been developed. In the mobile phases, an ion pair reagent, such as tertiary amines for anionic substances [50]. Metabolites of interest form a stable ion-pair with the reagent, which increases their retention. Although very powerful, this technique is not commonly used because the use of ion pairing reagents “contaminates” the MS for the opposite ion mode. For example, tertiary amines for retention of anionic substances are only used if the MS is operated in negative ionization mode, where they don’t interfere with the analysis. Switching to positive ionization mode would lead to a very high signal of these amines and ion-suppression.

For analysis of polar metabolites, Hydrophilic Liquid interaction Chromatography (HILIC) is employed. This method combines the polar stationary phases of NP with the solvents of RP and is well-suited for the analysis of amino acids, nucleotides, sugars, and other polar metabolites. Despite a large number of possible separation chemistry available, RP columns are used most. However, very polar metabolites are not retained on RPLC columns. Therefore, hydrophilic interaction liquid chromatography (HILIC) is frequently used. HILIC columns are either silica or derivatized silica columns [51, 52]. The different separation methods cover different parts of the metabolome. In order to achieve a thorough, comprehensive overview, several methods need to be combined.

Furthermore, different other stationary phases employing combinations of different retention mechanisms are available but not widespread in metabolomics, e.g. PFP. These columns can be used in different separation modes but is mostly used in RP mode. The PFP rings show a slight anion exchange characteristic, which leads to increased retention of polar metabolites, such as amino acids or organic acids [53]. A recent review investigation he uses of different chromatographic methods in metabolomics by curating information from the two public metabolomics repositories, MetaboLights and Metabolomics Workbench [54-56]. The analysis showed that RP is still the most prominent mode of analysis for metabolites and lipids.

Many metabolites carry (strong) ionic functional groups (such as carboxylic acid or phosphate groups) and are highly polar. These metabolites still often show broad and asymmetric peaks in HILIC. Capillary Electrophoresis (CE) coupled to MS is another powerful analytical technique in metabolomics [57]. In CE, charged molecules migrate in a liquid background electrolyte along an electrical field. Separation occurs due to different velocities of the metabolites, depending on the ion mobility and the electric field strength. However, CE-MS is often regarded as not as stable as LC-MS. This is particularly true for biofluids such as urine, which vary a lot in their salt content. Methods for the analysis of cations, anions, and nucleotides have been developed and applied to different matrices.

Although not as widespread as GC- or LC-MS, CE-MS offers several advantages, which makes this method an interesting alternative for metabolomics, especially if the sample amount or volume is limited. CE was originally adapted for Human Genome Project to replace slab sequencing gels. Coupling with MS requires dedicated ion sources, but the low flow rates and volumes make it ideal for hyphenated analysis, especially for low-volume samples. Since most metabolites contain at least one ionizable functional group, electrophoretic separation might be possible. Migration times are based on a metabolite's mass, shape, volume, and charge. Potential variation of migration times can be normalized by converting them to electrophoretic mobilities using migration marker substances [58, 59]. Analysis of the same samples using HILIC and CE has shown that the two methods provide orthogonal separation.

Lastly, ion mobility (IM) is another separation technique to enhance MS. However, in contrast to all previous techniques, it is used post-ionization and can also be combined with other separation techniques, yielding two-dimensional separations. IM describes ions traveling through a tube filled with gas along an electrical field. Ions are accelerated along the electrical field, and their velocity is reduced by collision with neutral gas molecules, e.g. nitrogen or helium. The amount of this reduction is dependent on the size and shape of the ion. Based on this different ions have different traveling velocities; hence they are separated based on their size and shape. Using IM, it is possible to separate isomeric and isobaric structures, which cannot be separated otherwise. The value measured by ion mobility is called collisional cross section (CCS), which refers to a molecule's rotationally averaged cross section. It is measured in Å² (angstrom square). This value is dependent on the adduct of a molecule; hence $[M+H]^+$ and $[M+Na]^+$ ions of the same metabolite have different CCS values. In contrast to retention time, which represents a system property defined by the analyte of interest, column, solvent, etc., the CCS value is a molecular property and, is therefore comparable between different instruments (within a certain error range).

1.2.5. Targeted and non-targeted metabolomics

Metabolomics experiments can be setup in two different ways. Although metabolomics was originally defined as a holistic approach, the different terminologies are often mixed or misused. To avoid confusion, the terms non-targeted and targeted metabolomics are used throughout this work. Non-targeted (or un-targeted) metabolomics aims to detect and (semi)quantify as many metabolites as possible in a given sample. No prior selection of metabolites or metabolite classes is made, nor does it require prior knowledge or hypothesis. This approach is typically used for exploratory studies and compares minimum of two sample sets, e.g. healthy and diseased against each other. Different methods are combined to achieve good coverage of the metabolome, e.g. HILIC-MS and RP-MS. Comparison between different metabolites is not directly possible due to differences in ionization behavior (e.g. phenylalanine and tyrosine are only separated by one reaction, but their ionization properties are quite different). Non-targeted metabolomic workflows include several steps like data processing, metabolite identification, and extensive statistical and bioinformatic analysis. Based on the obtained results, hypotheses on the involved metabolic pathways and their alteration are generated. These hypotheses should be followed up and validated by targeted metabolomics analysis.

In targeted metabolomics, several metabolites are preselected (e.g. from one or several pathways or a class of metabolites). Single methods are optimized toward the optimal detection and quantification of these selected metabolites. Typically, absolute concentrations of metabolites are obtained, which also allows the analysis of single samples. Furthermore, different metabolites can be compared due to absolute quantification. The selection of non-targeted or targeted metabolomics depends on the biological question.

1.2.6. Lipidomics

Lipidomics represents a sub-field of metabolomics, dealing with the comprehensive analysis of lipids. Lipids are hydrophobic or amphipathic molecules produced through thioester or isoprene unit condensations [60]. All lipids and related compounds within a biological model

form the lipidome. Lipids are believed to represent a relatively homogenous group of molecules having in common being soluble in organic solvents such as chloroform. However, they show an enormous diversity with many different side chain lengths and degrees of unsaturation possible, leading to an ample combinatorial space of structures.

Additionally, lipids span a wide range of concentrations [61]. Lipidomics has emerged in the past few decades as a promising field for developing applications in medicine and industry [62, 63]. Analysis of lipids has entered the main stage due to recent developments in MS technology, which made it possible to analyze a large number of lipids simultaneously. Two main approaches exist in lipidomics: shotgun and LC-MS-based lipidomics.

1.2.6.1. Shotgun lipidomics

Shotgun lipidomics directly infuses the raw lipid extract without prior chromatographic separation into an MS. The direct infusion enables continuous delivery of lipid samples, with a constant lipid concentration at all times. These constant conditions are beneficial for lipid quantification since ion suppression is constant and lipid analytes and internal standards are analyzed under identical analytical conditions. Lipid identification and quantification are performed on characteristic fragmentation patterns specific to lipid classes and species. If enough sample is available long infusion times can be used to increase the signal-to-noise ratio and enable, enough time for detailed tandem MS analysis is available. Due to missing chromatographic separation no isomers can be distinguished, and low abundant lipids arebe suppressed by higher abundant ones. However, the method is fast, highly reproducible, and has been employed even in large-scale cohorts [64]. Infusion can be performed using a gas-tight syringe and a syringe pump. However, this does not allow automation. Typically for automated infusion LC systems using the autosampler and pumps without a separation column or the Advion TriVersa NanoMate (a chip-based nanoESI device that is capable of spraying up to 1 hour from only a few µLs of lipid extract) are used.

Different shotgun lipidomics approaches have been developed. Tandem MS-based shotgun lipidomics uses neutral loss or product ion scans to detect specific lipid classes. This

methodology is often realized on low-resolution instrumentation such as triple quadrupole QqQ. Since lipids are filtered two times (in the first and third quadrupole), a high S/N can be achieved. Multidimensional shotgun lipidomics (MDSL) also uses DI-MS for lipid analysis [65]. However, differences in sample preparation allow for separating lipid classes upfront of analysis and combining different scanning techniques, such as neutral loss or precursor ion scans. The combination enables the identification of isomeric and isobaric lipid species still present within one extract.

Developing high-resolution hybrid instruments such as QqToF or Q-Orbitrap with increased duty cycles and higher resolving power enabled to reduce false positive identification in tandem MS analysis (e.g., the overlap of isobaric lipid species). Since from the obtained data, any neutral loss or product ions can be extracted from this type of analysis also allows retrospective interpretation for novel, previously unknown lipid classes or modifications. High-resolution MS enables the better separation of isobaric species leading to greater depth and more accurate quantification in lipid analysis [66]. Another possibility for the separation of isobaric lipid species is ion mobility [67].

1.2.6.2. LC-MS based lipidomics

LC-MS based lipidomics approaches are employed if separation of isomers and the analysis of low abundant lipids is required. Lipids are separated on a reversed-phase C8, C18, or C30 column using an ACN-iPrOH gradient, with variations in exact composition and/or used additives and modifiers [62, 68, 69]. Using RP, lipids are separated based on their hydrophobicity, which is determined by the length of acyls in the different lipids. Extensive coverage of different lipid classes is achieved, starting from polar lipids like lysophospholipids, fatty acids to non-polar lipids like triacylglycerols or cholesteryl ester. Along homologous series of lipids, typically linear trends are observed, which allows a more accurate lipid identification. Major drawbacks are the longer run-times (up to 30 minutes) and the elevate pressure caused by the use of iPrOH in the strong eluent. In contrast to RP, HILIC separates lipids based on their polar headgroup and class [70]. Since separation is

based on lipid class, internal standards are eluting with the analytes of interest, which in contrast, RP allows a more accurate quantification [71]. However, no separation of isomeric lipid species is achieved. Supercritical fluid chromatography (SFC) represents an interesting alternative to HILIC since separations are performed at higher efficiency leading to narrower peaks and short analysis times [72, 73].

1.3. *C. elegans* metabolomics & Lipidomics

Combining a genetically tractable and easily cultivatable model organism like C. elegans with global analytical approaches like metabolomics and lipidomics holds great promise to advance the knowledge of metabolic regulation, metabolism, and its role in health and disease. Several publications on *C. elegans* metabolomics and lipidomics have been published in the past years, and the number is steadily increasing. Both MS and NMR have been applied to *C. elegans* research question.

1D-^{1}HNMR is well suited for the analysis of highly concentrated metabolites, such as amino acids, organic acids, sugars, and others and has also been employed in *C. elegans* [74-78]. However, overlapping signals in 1D-^{1}H NMR results sometimes in ambiguous metabolite identification and can interfere with exact quantification. The use of 2D NMR in different ways helps improve the identification rates and has been used in C. elegans metabolomics [79-82]. In order to further enhance heteronuclear multidimensional NMR based on ^{13}C, labeling of *C. elegans* with ^{13}C can be performed to overcome the low natural abundance of ^{13}C. Labeled bacteria are used as food for the worm and can increase the sensitivity by two orders of magnitude, which enables the detection of much more metabolites than in one-dimensional NMR metabolomics [83, 84]. The major drawback of performing 2D ^{13}C HSQC experiments using ^{13}C labeled metabolites is additional structures arising in the spectrum from ^{13}C/^{13}C couplings, which are non-existent at natural ^{13}C abundance. These additional structures can reduce the possible sensitivity gain of ^{13}C labeling and increase the chances of peak overlap. Geier *et al*. investigated different HSQC pulse programs for fully ^{13}C labeled tissue extracts from *C. elegans* and found at constant time HSQC (ct-HSQC) leads to an improved peak shape and better peak detection of metabolites [85-89].

Differential analysis by 2D-NMR (DANS) has been developed using *C. elegans* and overlays and subtracts 2D-NMR spectra from two different conditions, enabling structural elucidation of differentially regulated compounds, even for minor components in complex matrices [90, 91].

As an alternative to the extraction of metabolites, direct analysis of nematodes using high-resolution magic-angle spinning (HR-MAS) NMR spectroscopy has been described [92, 93]. In this technique, the sample is spun at about 54.74° to the magnetic field to overcome magnetic field heterogeneities within solid samples responsible for broadening resonance lines. Though no extraction is required, around 1000 nematodes are necessary for standard HR-MAS analysis due to the general insensitivity of NMR. The usage of 1H NMR microprobes can enhance sensitivity, as demonstrated by Wong *et al*. for the metabolic phenotyping of a low number of *C. elegans* [94, 95]. In non-targeted metabolomics 1D-^{1}H NMR has been combined with other analytical platforms such as DI-MS, GC-FID, GC-MS, or LC-MS to increase the metabolite coverage [75, 76, 96, 97].

GC has been used early on in metabolomics, though not named like this at that time; for example Pauling *et al*. used GC for the analysis of urine and breath [98]. EI is typically used as an ionization method in GC-MS. EI yields highly reproducible but fragmentation-rich spectra. These spectra are useful for the dereplication of known and previously measured metabolites, and several EI spectra libraries are available. However, *m/z* of intact molecules is only occasionally observed, making unknown identification difficult. Chemical ionization (CI) as a soft ionization technique yields intact precursor ion *m/z,* which can be subjected to fragmentation in tandem MS. Therefore CI is preferred for analysis of unknown substances, as shown by Jaeger *et al*., who were using GC-APCI-MS for metabotyping in *C. elegans* [99].

Different MS tools were proven as effective tools for the analysis of the *C. elegans* metabolome, e.g. DI-MS for analysis of lipids [80, 100]. Lipid annotation was achieved by accurate mass and/or MS/MS characterization. Central carbon metabolites, such as organic acids or amino acids and fatty acids in *C. elegans,* have been analyzed by GC-MS [76, 81, 90, 96, 99, 101-105]. The latter, fatty acids, have also been analyzed by GC using a flame ionization detector (FID) to determine the fatty acid composition [75, 76, 96].

Like other metabolomics applications, MS is frequently coupled with UPLC to separate *C. elegans* metabolites before mass spectrometric analysis [82, 97, 103, 106]. RP is a dominant

separation mode for the analysis of non-polar metabolites and lipids. Lipids are typically analyzed on C18 column, but also C8 columns have been used [76, 96, 102, 106-109]. Analysis of polar metabolites is achieved by HILIC [51, 52] or ANP [110]. A less widespread method is ion-pairing RP in negative ionization mode, e.g. using tributylamine as ion-pairing reagent. [104]. Another possibility for analysis of polar metabolite is CE, which requires only a small sample amount, only a few nL per injection. Still, it has not been applied to *C. elegans* so far. However, its general usability in metabolomics has been reviewed extensively [111].

1.3.1. C. elegans metabolomics & lipidomics method development

Metabolite and lipid extraction is the first step in metabolomics and lipidomics. *C. elegans* represents an interesting case for sample preparation since it contains a hard cuticula. Geier *et al.* compared different homogenization and extract methods [82]. Six different tissue disruption techniques were combined with two solvent systems, 80% MeOH or a modified Bligh and Dyer extraction. Analysis was performed by GC-MS, LC-MS, and NMR. 80% MeOH in combination with a bead beater was found to represent the best solution for analysis of the polar metabolome. For NMR analysis high resolution magic angle spinning (HRMAS) represents an interesting alternative to laborious sample preparation. Blaise *et al.* used this method to study wildtype N2 and *sod-1(tm776)* [112]. Worms were washed off the plates and fixed with 3.7% formaldehyde. After extensive washing, including a final wash with D_2O. About 1000 worms were filled into an HR-MAS rotor.

C. elegans has a rich lipidome with several lipid classes present. Extraction of all lipids within a single method is complicated. Therefore, several extraction methods have been described. MTBE extraction has been developed using *C. elegans* samples and showed similar yields to chloroform-based methods [113]. The method is widespread and has been used in different studies, but it was also compared against different other extraction methods, such as Folch, Bligh & Dyer, or BUME [114]. Recently this method has been further developed to be combined with metal analysis from the aqueous phase [115]. However, several metabolites

or lipid classes might be of low abundance and require additional sample preparation steps such as preparative chromatography, TLC, or SPE. For example, using TLC, major lipid classes can be separated and further enhanced by using 2D-TLC, e.g. for the purification of glycolipids or other classes [80, 108]. Similarly, SPE based on anion exchange material has been used by Hänel et al. for the enrichment of sphingolipids [116].

1.3.2. C. elegans aging and longevity

C. elegans is a premier model system to study aging and longevity, with several genetic modifications increasing or decreasing lifespan [117]. Several metabolomics studies have been employed to study the effects of mutation of *daf-2*, which encodes for the orthologue of the insulin/insulin-like growth factor (IGF) receptor. Fuchs et al. used NMR-based metabolomics to study three different alleles of *daf-2* (e1370, m21 and m596) as well as *daf-28*(sa191) and *ife-2*(ok306) [118]. Additionally, dauer larvae have been included in the analysis. A common trend for all alleles of *daf-2* and *daf-28* (which disrupts a DAF-2 ligand) is the increase of branched-chain amino acids (BCAA) leucine, isoleucine, and valine. Out of all tested alleles, m41 showed the highest increase. Further analysis showed that these changes are dependent of the transcription factor DAF-16/FOXO, which acts downstream of DAF-2 and is activate upon *daf-2* mutation. *daf-2(e1370);daf-16(m26)* double mutants showed similar levels to wildtype or *daf-16(m26)* single mutants. Combination with transcriptomic analysis revealed several major shifts in carbohydrate and central carbon metabolism. Martin et al. similarly compared *daf-2(e1370)* and *daf-2(e1370);daf-16(m26)* mutants but additionally included mutation of *pept-1(lg601)*, which encodes for an intestinal di- and tripeptide transporter [74]. The additional mutation of *pept-1* even further enhances the lifespan of *daf-2* mutants. Analysis of *C. elegans* and their excreted exometabolome showed that worms carrying mutations of *pept-1* have a decreased one-carbon metabolism. Consistent with Fuchs *et al.* increased levels of BCAA were found. mRNA expression levels also showed increased activity of BCAA metabolic genes.

A time course analysis of metabolites across the age of wildtype and *daf-2(m41)* mutants was performed by Davies et al. using NMR [119]. Metabolite levels were correlated with age and several differences and patterns were detected in *daf-2(m41)* compared to wildtype worms. While several metabolites change in the wildtype worms and only show modest or no changes in *daf-2(m41)*, others show strong changes or even opposite responses. Similar to earlier studies BCAAs were found higher in young adult *daf-2(m41)* worms, but levels drop to the wildtype level at later stages. Another interesting trend was seen for trehalose, which is increased in *daf-2* mutants but not in the m41 allele in young adults, but shows a strong increase throughout aging also in *daf-2(m41)* mutants. Trehalose has been previously shown to increase the lifespan of wild-type worms but has not further extended the lifespan of *daf-2* mutant worms [120]. Lipid analysis showed an increase in triacylglycerols, especially those containing branched and monounsaturated fatty acids.

Castro *et al.* performed a metabolomic analysis of *daf-2* mutants using NMR, GC-MS and LC-MS [121]. NMR was used for the study of polar metabolites. GC-MS analysis was used for amino, and fatty acids and LC-MS was used for the analysis of lipids. Based on their obtained data, correlation analysis was performed. This analysis showed a correlation between the catabolism of BCAAs and e.g. 15-methyl hexadecanoic acid. Prasain *et al*. performed lipidomic analysis of *daf-1* and *daf-2* mutants using a workflow called $MSMS^{ALL}$ [122]. They identified changes in several phospholipid classes as well as di- and triacylglycerols. *daf-2* mutants are described to show a fat phenotype with a higher number of lipid droplets. This is mirrored by the lipidomic analysis showing that *daf-2(e1370)* has higher triacylglycerol levels than wildtype. Lourenço *et al*. investigated prohibitin deficiency in WT worms and *daf-2*(e1370) mutants [75]. They observed that prohibitin deficiency shortens the lifespan of wild-type nematodes while it significantly extends the lifespan of the already long-lived *daf-2* mutants. GC-FID analyses showed that prohibitin depletion of daf-2 results in changes of polyunsaturated fatty acid contents, which are already linked to longevity. Fatty acid composition and amino acid and carbohydrate metabolism, which was analyzed by

NMR, were more deeply affected by prohibitin depletion in wild-type nematodes compared to *daf-2* mutants.

Metabolomics is able to produce a static snapshot of the current metabolic state. However, in most cases, it remains elusive how this state was reached regarding which pathway was active. Using isotopic tracers can help capture the dynamic nature of metabolism. Tracer analysis in *C. elegans* represents a complicated task since no axenic medium is available, and always co-metabolism of tracers between *E. coli* and *C. elegans* is determined. Perez and van Gilst developed an interesting approach that allows following lipogenesis in *C. elegans* based on feeding a mixture of isotopically non-labeled or labeled *E. coli* [123]. The *E. coli* food source is grown in either non- or fully labeled growth medium. *C. elegans* was fed with a 50/50 mixture of non-labeled and labeled *E. coli*. Fatty acids from *C. elegans* have been analyzed with GC-MS. Fatty acids directly derived from *E. coli* are either 100% non-labeled or labeled, while all fatty acids that are produced by *C. elegans* show a mixed labeling pattern. Based on the degree of labeling, it can be determined if complete *de novo* synthesis or elongation has been performed. This method was applied to different alleles of *daf-2*: m577, e1368, m596, e1371, e1370, and m41. Interestingly, not all alleles showed the same extent in de novo fatty acid synthesis changes, with the highest levels found in *daf-2(1370)* and *daf-2(m41)*.

Although the exact way how changes in metabolism support longevity are not entirely understood, all results point toward a major rearrangement in the central metabolism. Depuydt *et al.* performed proteomic analysis of *daf-2(e1370)* mutants [124]. Consistent with metabolomics results, they have seen major changes in central metabolic pathways. Interestingly, the rewired metabolism is reminiscent of the metabolism in dauer larvae and points towards efficient use of internal nutrients.

Beside mutation in *daf-2* different other mutations can lead to an extension in lifespan. Butler *et al.* performed GC-MS-based analysis of the exometabolome of mutants in the mitochondrial electron transport chain [125]. An increase of branched-chain α-keto acids and

their reduction products was seen explicitly in mutants of mitochondrial electron transport chain genes but not in *daf-2(e1370)*, *clk-2(qm37)*, *eat-2(ad465),* and *slcf-1(tm2258)*, which are also long-lived. Furthermore, their profile is different from worms cultured under anaerobic conditions. Further analysis showed that inhibition of dihydrolipoamide dehydrogenase resembles the observed metabolic phenotypes.

One of the metabolomics's significant advantages, but also obstacles, is that metabolism is one of the first things to react upon a stimulus. Therefore, great care needs to be taken when samples are taken and processed. Furthermore, chemicals used e.g. during cultivation, might lead to unexpected results. 5-Fluoro-2-deoxyuridine (FUdR) is used in *C. elegans* cultivation to maintain synchronous culture [126]. It inhibits DNA synthesis and the production of offspring. Davies *et al.* evaluated the effect of FUdR on the metabolic response in *daf-2* mutant worms [127]. *C. elegans* were cultivated with or without FUdR and metabolites were analyzed via NMR. Interestingly, PCA separated worms grown with or without FUdR on the first principal components. Only if the dataset was divided into two separate sets with or without FUdR a separation between wildtype and *daf-2* mutants was visible. Treatment with FUdR let to different changes between wildtype and *daf-2* mutant worms, e.g. for glutamate.

Different signatures in changes might be shared between other long-lived mutants, for example, *daf-2* and *eat-2* mutants. The latter is used as a model for dietary restriction (DR). Gao et al. compared both mutants using metabolomics and transcriptomics [101]. Both mutants show an increase of glycerolipid and purine degradation intermediates and AMP, while amino acid levels and certain fatty acids decreased. Longevity signatures in the transcriptome and metabolome observed in both mutants show an increase in amino acid metabolism, potential due to lower protein synthesis and an up-regulation of purine metabolism. Another model for longevity are dauer larvae, which can live about eight times longer than WT worms, and changes in metabolism can be studied using metabolomics [77, 128]. Significant differences to non-dauer worms, but similar changes to long-lived mutants

were found. Comparable to other studies changes, dauer larvae contain metabolites of distinct central carbon metabolic pathways such as carbohydrate, amino acid, and choline metabolism.

Sirtuins are NAD+-dependent protein deacetylases, and mutation of the *sir-2.1* gene in *C. elegans* also leads to a lifespan extension of up to 50% [83]. Using ^{13}C-HMN-NMR comparing *sir-2.1* mutant and WT *C. elegans* revealed differences in BCAAs, triacylglycerols, carnitine and acetyl-CoA, all elevated in WT worms. In contrast, lactate, glutamate, alanine, fatty acids, and AMP were increased in the mutant worms pointing to changes in glycolysis, nitrogen catabolism, and lipolysis.

Mutation in the *glp-1* gene leads to the ablation of germline stem cells and, therefore, infertility, but also an extension of lifespan [129]. *glp-1* mutants show an altered lipid metabolism, and many age-related metabolites, including increased levels of pyrimidine and purine metabolism intermediates and decreased concentration of citric acid cycle metabolites, were also differentially regulated in *glp-1* mutants worms [97]. NMR and UPLC-MS showed that *glp-1* mutant worms show similar changes as WT during ageing. However, dysregulation of some metabolites (e.g. GSSG, valine, leucine, malate and serine) was observed, while others (e.g. cystathionine, glycine, arginine and trehalose) didn't change. In order to pinpoint common metabolic signatures in aging and longevity in the future more meta-analysis of the previously collected data is required.

Elevated trehalose has been observed in almost all long-lived *C. elegans* mutants. It was therefore tested if supplementation with trehalose positively affects the lifespan of *C. elegans* [119]. Such a positive effect was only observed for middle-aged WT worms treated with trehalose but not for early adults. Comparing worms at different ages (10-day adults vs. young adults), differences in reduced and oxidized glutathione were observed. While the first one was reduced, the second one increased. The antioxidants taurine and hypotaurine declined with age. Phosphocholine and trehalose showed higher concentrations in aged worms, while purine metabolism intermediates decreased. Intermediates of pyrimidine

metabolism and TCA cycle were variable, with no apparent pattern. This is in contrast to short-lived mutants such as *mev-1(kn1)* studied by Jaeger *et al.* with GC-EI- and GC-APCI-MS [99]. The found an upset TCA cycle balance, elevated lactic acid fermentation, and increased amino acid catabolism. The effect of aging was also studied by HR-MAS-NMR comparing young adults and seven days old adults [93]. While saturated and unsaturated lipids, glycerophosphocholine, phosphocholine, glutamine, and glycine increased with age, several amino acids such as alanine, arginine, leucine, isoleucine, lysine, phenylalanine, tyrosine, valine and glutamate and other metabolites such as acetate, lactate, glycerol, formiate, and cystathionine decreased.

Analysis of age-related metabolites has been performed using HILIC-UPLC-MS combined with flux balance analysis (FBA) with a new approach called MetaboFBA [52]. Changes similar to earlier publications have been observed with levels of some amino acids (serine, threonine, leucine, lysine, glutamine/ glutamic acid, methionine, tryptophan, arginine, homoserine, cystathionine) and nucleotides (guanosine, cytidine, uridine, GMP) decreased over time, while guanosine, cytidine, uridine, GMP decreased. Betaine, carnitine, leucic acid, pantothenate, kynurenic acid, xanthurenic acid, and degradation products of both nucleotides hypoxanthine, xanthosine, and allantoin increased over time.

Changes in metabolism have not only been studied during aging but also development. Analysis of fatty acids using DI-MS and targeted metabolomics using HPLC-MS has been performed in embryos and L1 to L4 larvae [130]. Fatty acids showed the lowest abundance in the larval stages and increased in adult worms. Amino acids peaked in L3 and in early adults and decreased throughout aging. The only exception was asparagine, which showed the highest levels in eggs and larvae but then decreased. Phosphatidylcholines (PCs) and phosphatidylethanolamines (PEs) were present in higher abundance in larval stages and young adults and decreased during the adult lifespan. Lysophosphatidylethanolamines (LPEs) showed a similar trend, while cardiolipins (CLs) were low in L2 and increased during

early adulthood. Phosphatidylglycerols (PGs) and sphingomyelins (SMs) showed an opposite pattern.

The nuclear hormone receptor DAF-12 regulates both adult lifespan and larval development. Using 2D-NMR and DANS between *daf-22* and *daf-9;daf-12* mutants have been combined with activity-guided fractionation to identify DAF-12 ligands, including Δ4- and Δ7-dafachronic acid already known, but also new derivatives [90]. Separation and independent quantification of both isomers are important since they are derived from different branches of a pathway involving different enzymes [131]. A method based on a 1.3 µm particle column, showing high efficiency, was developed to separate the two isomers and quantify them in dietary-restricted worms [132, 133].

Wang *et al*. used UPLC-MS and GC-MS to find metabolic changes in *pnc-1* mutants that are deficient in nicotinamidase and show slow reproductive development [104]. The targeted analysis identified NAD+ and glycolytic intermediate levels suggesting that efficient glycolysis seems to be mandatory for reproductive growth.

1.3.3. C. elegans ascaroside signaling

Individual *C. elegans* communicate with each other using a blend of small molecules excreted into the environment. The best-known example is ascarosides, which constitute the dauer pheromone. Chemically, they have been identified as glycosides of the di-deoxy-sugar ascarylose. They regulate development and the (coordinated) behavior of the nematode, such as dauer formation, male or male or hermaphrodite attraction or repulsion, or aggregation. Ascarosides contain a hydrophobic tail derived from long and very long chain fatty acids and can contain several modifications at various positions. Metabolomics approaches have been used to identify different members of this class. Srinivasan *et al.* used activity-guided fractionation, 2D-NMR, and targeted LC-MS measurements to identify mating signals, which were excreted by L4, YA, and adult hermaphrodites [79]. Three ascarosides called ascr#2, ascr#3 and ascr#4 were identified, and the attraction seemed to be

concentration-dependent. While males were attracted at low concentrations, deterrence occurred at higher concentrations.

Further investigations using differential analysis by 2D-NMR in combination with LC-MS led to the discovery of previously unknown ascaroside species called ascr#6.1, ascr#6.2, ascr#7, and ascr#8 [91]. It has been shown that DANS is well suitable for discovering ascarosides since it can resolve them from a chemical background in the spent cultivation media such as amino acids, peptides, and other compounds that otherwise dominate the total ion chromatogram (TIC) at LC-MS. However, LC-MS was more beneficial for the detection of low-concentrated ascarosides as ascr#1, ascr#6.1, and ascr#6.2.

Pungaliya *et al.* employed differential 2D-NMR to the exometabolome of N2 wildtype worms and *daf-22* mutants and identified three known and four unknown ascarosides [134]. One of these molecules contained an unusual p-aminobenzoic acid moiety. Interestingly, this residue seems to be specifically attached to ascr#7, which features a seven carbon long side chain, although the nine carbon side chain member acsr#3 is found in higher amounts.

A similar 2D-NMR approach was used by Izrayelit *et al.* to identify the interaction between the ascarosides and another class of bioactive molecules endocannabinoids [135]. Analysis of the daf-22 exometabolome revealed the presence of several ethanolamide-modified ascarosides. These molecules are longer compared to commonly found ascarosides with chain lengths between 21 and 29 carbons. HPLC-MS analysis confirmed the presence of long-chain ascarosides and ethanolamides. Von Reuss *et al.* identified several other modified ascarosides using LC-MS/MS-based comparative metabolomics [136]. Ascarosides are putatively synthesized by peroxisomal β-oxidation of long-chain precursors. The enzymes ACOX-1, MAOC-1, DHS-28 and DAF-22 are required for their correct biosynthesis. Based on the LC-MS/MS analysis, they profiled different ascarosides, including saturated, α,β-unsaturated and β-hydroxylated species in mutants of the above-mentioned enzymes. Results proved the suggestion that ascarosides are produced by β-oxidation since particular profiles were obtained for the respective mutants. In addition to the normally found ω-1

hydroxylated side chains, molecules also containing ω-hydroxylated species have been identified, together with indole, *p*-hydroxybenzoic, or tiglic acid-modified molecules. Indole-containing ascarosides serve as potent attraction and aggregation signals for hermaphrodites. By using isotope labeling experiments, it was shown that the indole moieties derived from Tryptophan are produced by *C. elegans*. Again, the indole containing ascarosides show a very specific chain distribution, which was the basis for the investigation of different proposals for biosynthetic pathways [137, 138]. One of the largest analyses of ascarosides was performed by von Reuss et al. using GC-MS [139]. Using a specific ascarylose-derived fragment together with fragments specific to the side chain, they were able to identify about 200 ascarosides. These included previously known ω-1 linked acyl saturated, α,β-unsaturated, β-hydroxylated and β-keto derivates together with ω-linked, ethanolamides, and others. Parallel to GC-MS, LC-MS/MS analysis has been applied for further confirmation. Also, applying LC-MS/MS Artyukhin *et al.* used molecular network analysis to study the metabolomic "dark matter" [140]. Many previously unknown structures have been detected by performing analysis in positive and negative ionization modes. Interestingly, some new molecules related to the ethanolamide containing ascarosides have been detected, including N-acylethanolamides, glycerophosphoethanolamides, and glycerphosphoethanolamides glucosides.

1.3.4. C. elegans lipidomics

C. elegans is a premier model organism for the study of lipid metabolism. Regulation of fatty acid and lipid metabolism has been extensively studied in *C. elegans* [141]. The transparent body makes it easy to follow fat storages or the expression of lipid-related genes using reporter constructs. Additional techniques and different to analysis methods for polar metabolites are necessary due to the very different physicochemical properties of lipids. Various possibilities for the analysis of fatty acids and lipids exist and have been applied to *C. elegans*. GC-FID or GC-MS are most widely used to analyze fatty acids and sterols, while intact lipids are analyzed through shotgun or LC-MS-based lipidomics. Methods for the analysis of lipids and lipid metabolism are reviewed elsewhere [142]. Analysis of fatty acids

has been performed by many research groups using GC-FID or GC-MS. The genetic basis for the biosynthesis of fatty acids has been unraveled by Watts and Browse using GC-FID [143]. Another technique that has been used quite extensively is thin-layer chromatography (TLC) [144]. Analysis of the total or lipid class-specific fatty acid composition of *C. elegans* has been carried out several times [76, 105, 145-151]. Most recently, Henry *et al*. determined the fatty acid composition of WT worms using EI and CI at GC-QToF-MS utilizing the high separation power of GC with high resolution of TOF-MS identify in total of 28 fatty acids based on exact mass and isotope pattern in CI [105]. Fragmentation in EI was used for further verification and identification. However, the actual fatty acid content is highly dependent on culture conditions and the bacterial diet [147]. The nematode is able to synthesize all required fatty acids, including monounsaturated (MUFA), polyunsaturated (PUFA), and branched-chain fatty acids (BCFA) on its own, although the majority of saturated fatty acids are directly taken up from the food. Interestingly, mono-methyl BCFAs (mmBCFAs) are produced by 99.9% by the worm itself [152]. While the *E. coli* diet of *C. elegans* is rich in the saturated fatty acids lauric acid, myristic acid and palmitic acid, they are only minor species in *C. elegans*. Interestingly, palmitic acid and stearic acid have high percentages in phosphatidylinositols (PIs) [151]. The cyclopropane fatty acids cis-9,10-methylene hexadecanoic acid (FA 17:0[9-10cy3:0]) and cis-11,12-methylene octadecanoic acid (FA 19:0[11-12cy3:0]) are taken up exclusively from the bacterial diet and are enriched in triacylglycerols. Castro *et al*. investigated the regulation of the lipid content in *fat-5*, *fat-6,* and *fat-7* mutants [76]. They used GC-FID to analyze fatty acid profiles and LC-MS for the analysis of intact lipids. Common to most of the studied mutants was an increase in triacylglycerols and a decrease in phosphatidylcholines with unsaturated fatty acids, consistent with the role of these genes in the biosynthesis of PUFAs.

Despite the long history of *C. elegans* lipid analysis, the analysis of intact lipids and complete lipidomes only recently entered the *C. elegans* toolbox. One of the first analyses of intact phospholipids in *C. elegans* was conducted by Ishida *et al.* using nano Electrospray-Fourirer Transform Ion Cyclotron Resonance Mass Spectrometry (nanoESI-FT-ICR-MS) [153].

Utilizing the ultrahigh resolution of FT-ICR-MS, they were able to distinguish between diacyl and alkyl-acyl species solely on mass. Different other groups used direct infusion or shotgun lipidomics approaches to study the *C. elegans* lipidome. Schwudke et al. performed a lipid analysis on *C. elegans* undergoing the knockdown of two putative methyltransferases, PMT-1 and PMT-2 [100]. Knockdown of the two genes showed the increase of either monomethyl-PEs or dimethyl-PEs, proofing the putative gene annotation and their involvement in the biosynthesis of PCs from PEs in *C. elegans*. *pmt-1* and *pmt-2* are homologs to plant methyltransferase. *pmt-1* deficient worms arrest their development at the L4/early adult stage and *pmt-2* worms at the L3 stage. Upon silencing of *pmt-1* a decrease in MMPE and DMPE was observed, while RNAi of *pmt-2* led to an increase of only MMPE.

Lipid extraction is a crucial first step in the analysis of lipids. The extraction protocols according to Folch or Bligh and Dyer are the gold standards for lipid extraction. However, the lipid-containing organic chloroform phase forms the lower phase, with the interphase containing DNA, RNA, and proteins floating on it. This makes the recovery of the lower phase complicated. Matyash *et al.* developed a protocol using MTBE, which shows a lower density than water and forms the upper phase [113]. *C. elegans* eggs were used as a sample for the comparison of the Folch and the MTBE extraction protocols. Shotgun lipidomics using a qToF-MS was used for analysis. Different lipids from several lipid classes have been detected. The results showed that the two extraction methods have comparable yields.

Based on TLC and shotgun lipidomics, Penkov *et al.* identified maradolipids as a new lipid class exclusively found in *C. elegans* dauer larvae [80]. 2D-TLC was performed for lipid class separation, and spraying with Molisch reagent identified a class of glycolipids in dauer larvae. Chemical analysis with MS and NMR identified these lipids as 6,6'-diacyl trehaloses. Papan *et al.* performed shotgun lipidomics comparing L3 with dauer larvae and identified lyso-maradolipids, potentially representing a biosynthetic intermediate or degradation product of maradolipids [154]. Recently this class has also been investigated using UHPLC-IM-TOF-MS [155].

Sphingolipids are an important class in *C. elegans*. The nematode produces C17iso branched chain sphingoid bases using FA 14:0(13Me) instead of palmitic acid. Different studies used lipidomics to study ceramides and other sphingolipids. Ceramides are produced by three different ceramide synthases. One of these, HYL-2, has been shown to be protective under anoxic conditions [156]. The two enzymes HYL-1 and HYL-2 have different substrate specificity in regard to N-Acyl bound fatty acids. A further study showed that HYL-1 is required for the synthesis of sphingolipids with ≥ C24, while HYL-2 is for the synthesis of ceramides with ≤ C22 [157].

The fatty acid profile of sphingolipids shows a different fatty acid profile compared to glycero- and glycerophospholipids. Typically, mild alkaline saponification is used to generate free fatty acids from glycero- and glycerophospholipids, which are in turn, analyzed as their methyl esters by GC-MS. Amide-bound fatty acyls in sphingolipids are inert under this condition and require a different strategy. Therefore, total fatty acid profiles normally represent ester-bound fatty acids ignoring fatty acyls from sphingolipids. Chitwood *et al*. and Gerdt *et al*. performed an analysis of glucosylceramides in *C. elegans* and identified several 2-hydroxy fatty acids, which are not found in glycero- and glycerophospholipids. These fatty acids represent long-chain saturated fatty acids ranging from 16 to 26 carbons and include odd-numbered chains. Interestingly, also even numbered mmBCFAs were reported [158, 159]. Recently, UPLC-UHR-TOF-MS in combination with extensive fractionation of the lipidome, has been used to study sphingolipids in more detail [116]. Besides different ceramides, glycosylceramides, and sphingomyelins, unusual PE or mono-methyl-PE-modified glucosyl ceramides have been identified in a screen searching for molecules important under sterol deprivation [160]. Structural elucidation of such novel lipids is an important topic. Zhao *et al.* studied the fragmentation of SM species in *C. elegans* detected as $[M+HCO_3]^-$ adduct, which led to a radical-directed dissociation for detailed elucidation of N-acyl chains and sphingoid bases [161].

Another specificity in the *C. elegans* is the high amount of plasmenyl-PE (PE-P) and plasmanyl-PE (PE-O), with 5.1 and 4.9% of the phospholipid content, respectively. In mammals, ether lipids normally belong to PCs. In *C. elegans,* PE ether lipids mostly contained an 18:0 side chain at the sn1 position linked as ether or vinyl-ether [162]. The biosynthesis of ether lipids was investigated by Shi *et al*. [149]. Key enzymatic gene mutation leads to an increase in *de novo* fatty acid biosynthesis and reduction in the desaturases *fat-5* and *fat-7*, which potentially represents an orchestrated response to remodeling the lipidome composition.

Different treatments lead to profound changes in the lipidome, dietary restrictions (DR), for example. Klapper *et al*. studied additional supplementation with choline during DR and they have seen that lipid stores are utilized at a later time point [108]. While choline supplementation changed the lipid composition, it was not able to reverse the lifespan increase under DR. Excessive energy in *C. elegans* is stored in triacylglycerols, the primary lipid class found in lipid droplets, energy storage organelles. Typically, triacylglycerols are believed to be relatively inert and are only utilized if energy is required. However, the can serve as a buffer and reserve for specific fatty acids or can store cytotoxic fatty acids, e.g. large amounts of saturated fatty acids. Embryos rely on the energy stored in yolk during their development. A genetic screen searching for large lipid droplet phenotype in embryos was performed by Schmökel *et al.* [109].

Interestingly, *asm-3*, a member of the acid sphingomyelinase gene family, was identified as an interesting candidate. Interestingly, the large lipid droplet phenotypes were specific to *asm-3,* and not effect was observed in the closely related genes *asm-1* and *asm-2*. Performing lipid analysis by LC-MS/MS, changes in particular lipid species were found, while the total lipid content and the relative class composition didn’t change extensively.

Several genes coordinate lipid metabolism and have important regulatory functions. A vital regulator is the transcription factor SREBP-1 (sterol regulatory element binding protein), which has important role in lipogenesis. For example, low levels of PCs stimulate SBP-

1/SREBP-1. Smulan et al. identified *lpin-1* and *arf-1.2* as essential genes for the activation of SBP/SREBP-1 under low PC conditions [163]. Another important is *lipl-5*, a lipase-like gene, important during starvation. Macedo *et al*. studied mutants lacking LIPL-5 under *ad libitum* or starvation conditions by LC-MS analysis [164]. Four hundred nine different lipid species were detected, and differences in the activity of mitochondria accompanied the most prominent differences in detected cardiolipins. Generally, there is an intimate link between mitochondrial activity, lipid metabolism, and longevity. Mitochondrial dynamics can influence lipid profiles [165]. *fzo-1* mutants showed fragmented mitochondria, decreased mitochondrial membrane potential, and induced mitochondrial unfolded protein response. The lipidome analysis by UPLC-UHR-ToF-MS revealed a specific signature of changes in TGs, with fewer carbons and double bonds down-regulated in *fzo-1* mutants, while longer and more unsaturated TG species were upregulated. A further important transcription factor is HLH-30 linking mitochondrial dynamics and translation to longevity [166]. Furthermore, HLH-30 is central to the response to starvation, as *hlh-30* mutants are susceptible to starvation. Metabolite and lipid profiles of wildtype and *hlh-30* mutants have been compared over starvation and changes in long-chain acyl-carnitines and cardiolipins with an increase over the starvation time course. In contrast, *hlh-30* mutants showed no change [167]. Using assays to measure mitochondrial β-oxidation a shift towards peroxisomal β-oxidation was observed in *hlh-30* mutant. The additional knockdown of *prx-5* leads to hypersensitivity of *hlh-30* mutants to starvation. *prx-5* mutants have also been studied, and changes in TGs and ether PEs were observed [168].

The role of lipids in aging has also been studied. Combinations of different drugs, which extend the lifespan of the nematodes, were used, and the lipid profiles were compared by LC-MS/MS [169, 170]. The effects of the drugs depended on the presence of SBP-1/SREBP-1, and changes in the MUFA to PUFA ratio have been linked to this extension. Generally, the regulation of PUFAs is important in many aspects of *C. elegans* biology. Polar metabolites and lipids have been studied over a period of ten days under different feeding conditions [130]. Changes in fatty acids and phospholipids were regulated by MDT-15, which controls

the expression of *fat-7,* and differences were identified on the fatty acid and lipid species levels. In a similar fashion, changes in the lipid metabolism of dauer larvae were studied by LC-MS/MS [171]. Dauer larvae membranes are enriched in PUFAs, and the release of PUFAs and production of downstream metabolites such as eicosanoids are observed upon the end of the dauer stage.

Perez and van Gilst used GC-MS to study differences in fatty lipogenesis between different alleles of *daf-2* mutation [152]. The same approach was used to further study membrane dynamics [172]. Comparing the percentage of newly incorporated fatty acids into either phospholipids or neutral lipids, they showed higher turnover rates for phospholipids. However, the amount of fatty acids from *de novo* synthesis was only for palmitic acid significantly different. Turnover rates were determined in different mutants and identified *fat-5*, *fat-6* and *fat-8* as membrane maintenance regulators. Furthermore, ^{15}N tracers were used to follow the phospholipid turnover in more detail. This analysis revealed that *fat-7* also influences the turnover of lipids towards slower metabolism.

1.3.5. Nutrition and other topics

The metabolome of *C. elegans* is highly dependent on the food source and nutritional status [42]. One central point influenced by the bacteria strain used for feeding is the fatty acid composition [76]. Higher levels of the odd chain fatty acids of lengths 15 and 17 were found when *C. elegans* was fed with *Bacillus subtilis* instead of *E. coli*. A large portion of the fatty acid composition reflects the bacterial diet, but several fatty acids are also absent in bacteria and are therefore synthesized *de novo* in the worm. In contrast, levels of amino acids are almost identical between the different feeding conditions, suggesting that amino acid metabolism and their levels are tightly controlled and regulated.

Besides the topics covered so far, several others have been studied in *C. elegans* using metabolomics. For example, Hughes et al. studied the response of *C. elegans* to exposure of cadmium using LC-MS [173]. Different genotypes, WT, and *pcs-1* mutants, *mtl-1;mtl-2* double mutants and *mtl-1;mtl-2;pcs-1* mutants were used and decreased levels of

cystathionine, and increased levels of phytochelatins were found. This is due to the upregulation of the methionine transsulfuration pathways as a response to cadmium exposure. Likewise, the metabolic response to nickel and chlorpyrifos has been investigated and similar changes have been found for both exposures, suggesting a common detoxification response [174]. Schlipalius *et al*. investigated the metabolome of WT and *dld-1* mutants after exposure to phosphine [78]. Lastly, Sudama *et al*. investigated changes in metabolites due to lead exposure, but instead of LC-MS, they employed HPLC with a coulometric array to find after lead exposure, changes in tryptophan, tyrosine, and purine were observed in WT worms [175].

Different transgenic models of diseases are possible in *C. elegans*. One of them is Alzheimer's disease. Van Assche *et al.* studied changes in the metabolism after induced expression of the human amyloid-β peptide by GC-MS and LC-MS using RP and aqueous normal phase (ANP) -MS separation [176]. Previous results from experiments in alzheimers disease were also obtained in humans, as, for instance, increased levels of allantoin could be confirmed [110]. Moreover, Teo *et al*. combined metabolomics data with transcriptomics and computational modeling and linked amyloid-β expression to TCA cycle [102]. The fast development and reproduction of *C. elegans* make it also an ideal organism to study the accumulation of mutations. The effect of the mutation accumulation has been studied by GC-MS and 29 metabolites have been identified, which vary in their vulnerability to mutation [177]. Additionally, new metabolite classes are identified on a routine basis in *C. elegans*, beside ascarosides. One example are anthranilic glucosyl esters, which show a blue fluorescence and accumulate during cell death [178].

1.4. Embedding *C. elegans* metabolomics and lipidomics in systems biology

C. elegans integrates several features that make it an interesting organism for systems biology approaches, including different forward and reverse genetic tools and the possibility to grow a large number of worms in a short time. Integrating metabolomics and lipidomics into the systems biology landscape requires that many samples and/or conditions can be measured with experimental setups, e.g., in genome-wide or drug screens. This translates into an increased number of samples, but a decrease in sample amount or volume per sample, since standard growth conditions are often hard to parallelize to reach the required throughput. While it is relatively easy to cultivate several thousands of worms for an experiment with a few single genotypes or conditions, genetic screens in *C. elegans* are typically carried out in 96-well plate format. Under these conditions, 10-15 worms can be cultivated in a single well without overcrowding. This means, on average, about a 100-fold reduction of biological starting material compared to standard culture conditions. So far, most of the *C. elegans* metabolomics and lipidomics studies used standard culture conditions of worms with several thousand individual worms per sample or even liquid cultures to obtain even higher numbers. Typical numbers from selected publications are summarized in Table 3. Depending on the metabolites of interest and their abundance, the number of required worms and culture plates must be increased and potentially pooled for a single replicate, lowering throughput. No direct correlation between the number of employed worms and reported metabolites exists. The detection of highly abundant lipids, for example, is possible with about 500 worms, while the detection of dafachronic acids, for example, requires larger amounts (> 10.000 worms) [132].

It can be concluded that either a reduction in biological information obtained from metabolomics and/or lipidomics experiments is expected or that more sensitive appraoches must be employed. Besides the required increase in sensitivity, the time needed for the analysis becomes another critical issue. Using longer runtimes (e.g. 15-30 minutes or above) in such large-scale experiments is unrealistic, which might introduce additional unwanted variation, either on the long storage time of samples or on more significant differences in

growth conditions along the time axis (e.g. different batches of cultivation media, etc.). A further important aspect is biological variation. Using standard cultivation techniques, variation between animals is averaged out on the population level, but if the number of animals is reduced, variation is potentially increased.

Table 3: Comparison of number of worms used for metabolomics and lipidomics analysis against the number of reported metabolites and the used analytical technique for publications indicating the number of used worms. The number of worms is very heterogenous as well as the number of metabolites reported. This number does not reflect the number of detected features, but only the number of metabolites reported in figures, tables and supplementary information. Furthermore, the abundance of the metabolites of interest plays an important role determining the required amount of worms. Lipids can be easily analyzed from a lower number of worms, while for example detection of dafachronic acids requires several hundreds of thousands.

Publication	Amount of worms / per sample	Number of metabolites reported	Technique
Blaise *et al.*, 2007 [179]	unknown	58	NMR
Atherton *et al.*, 2008 [180]	2000 worms	77	NMR / GC-MS
Hughes *et al.*, 2009 [181]	8000 worms / sample	10	NMR / LC-MS
Jones *et al.*, 2011 [182]	2000 worms / sample	65	NMR / GC-MS
Lucanic *et al.*, 2011 [101]	~ 5 x 10e6 worms / sample	6	GC-MS
Butler *et al.*, 2012 [183]	80-100,000 worms / sample	66	GC-MS
Butler *et al.*, 2013 [125]	120,000 worms /sample	69	GC-MS
Wong *et al.*, 2014 [94]	1000 worms / sample down to 1 worm / sample	49	NMR
Jaeger *et al.*, 2014 [99]	500 worms / samples	113	GC-MS
Gao *et al.*, 2017 [130]	2000 worms / sample	64	LC-MS
Gao *et al.*, 2018 [51]	2000 worms / sample	104	LC-MS
Folick *et al.*, 2015 [184]	200,000 worms / sample	71	LC-MS / GC-MS
Schmökel *et al.*, 2016 [109]	100.000 eggs	81	LC-MS
Schwudke *et al.*, 2007 [100]	12,000 worms	164	Other
Mahanti *et al.*, 2014 [90]	540,000 worms (estimated from 540 10 cm plates)	6	NMR
Mosbech *et al.*, 2013 [157]	1000 worms	47	Other

Publication	Amount of worms / per sample	Number of metabolites reported	Technique
Hastings *et al*. 2019 [52]	2000 worms / sample	104	LC-MS
Witting *et al.*, 2015 [185]	1000 worms / sample	2	Other
Witting *et al.*, 2014 [68]	5000 worms / sample		LC-MS
Witting *et al.*, 2017 [186]	5000 worms / sample		LC-MS
Witting *et al.*, 2015 [132]	20,000 worms / sample		LC-MS
Thondamal *et al.*, 2014 [133]	20,000 worms / sample	2	LC-MS
Papan *et al.*, 2014 [154]	30000 worms / sample	16	Other
Dall *et al.*, 2021 [187]	500 worms / sample	1244	LC-MS
Häußler *et al.*, 2020 [165]	500 worms / sample	540	LC-MS
Rackles *et al.*, 2021 [168]	500 worms / sample	844	LC-MS

The increase in sensitivity of analytical instrumentation allows the decrease of the required amount of biological starting material. A clear trend towards a lower number of worms is visible for standard metabolomics applications. The first steps in this regard have been performed by the use of DI-FT-ICR-MS, for example [185]. Metabolite extracts of 1000 worms have been produced, with an approximate concentration of 1000 worms / mL, but for optimal data acquisition, this extract had to be diluted 1:50, yielding a sample of 20 worms / mL. From this extract, only a few µl have been used for direct infusion, translating to the equivalent of less than a worm used. This result encourages that it is possible to perform metabolomics and lipidomics on a small amount of biological starting material. In extreme cases, analysis of the metabolome and/or lipidome of single worms might become possible in the future. To analyze the limited protein/metabolite amount within one worm, technical issues have to be overcome to enable single-worm proteomics or metabolomics. A first glimpse of single-worm metabolomics was published in 2014. Wong and co-workers employed a µHR-MAS-NMR system for the analysis of metabolites from intact *C. elegans*

[94]. They could identify 31 metabolites from only 12 animals. However, the measurement time was 2.5 hours for one sample. Using just a single worm, signals from choline moieties, glucose, tyrosine, and phenylalanine could be detected, but 15 hours of the acquisition were required. More recently, in a groundbreaking paper, Bensaddek *et al.* described the proteome analysis from single worms using a tailored microproteomics method [188]. Using this method, they were able to detect ~ 3000 proteins from a single *C. elegans* compared to ~5000 proteins from a standard macro-proteomics approach employing ~40,000 pooled animals. This new approach is powerful compared to previous microscopic studies using only a limited number of reporter systems for single proteins, while proteomics is able to analyze thousands of proteins in parallel.

These obtained results indicate that the inclusion of metabolomics and lipidomics in large-scale *C. elegans* experiments such as genome-wide screens, will be possible. This will generate new insights into the role of metabolism, its regulation, and implications for different aspects of the nematode's biology. However, various elements need to be developed better before entering this stage of analytical method development and applications.

1.5. Goals of this work

Metabolites show a large diversity in structure, and no unifying physicochemical principle like for the analysis of DNA and RNA exists. Therefore it is advantageous to first identify and define the composition of the metabolome and lipidome of the organism to study (even if it is only a draft). Such a collection of metabolites and lipids enables the estimation and to project into required methods for analysis. Likewise, methods for bioinformatic analysis are necessary to be able to cope with the large scale of obtained data.

The goal of the culminated work presented in this work is to develop further metabolomics and lipidomics approaches and data analysis strategies that allow delving deeper into the metabolism of *C. elegans*. Here the first goal it to better understand the complexity of the metabolic network of the nematode. No *C. elegans*-specific metabolome or lipidome database currently exists, though several genome-scale metabolic models have been described. However, scattered efforts lead to suboptimal performance and incompatibilities. Therefore, metabolic models were integrated into a single model, which serves as the basis for a *C. elegans* metabolic pathway database, including metabolite and lipid structures. Genomics, transcriptomics, and proteomics benefited from joint efforts in creating community-accepted standards. Some effort is also made in *C. elegans* towards harmonized metabolomics, but a central resource is required to foster further standardization and comparability. Since only limited material will be available in large-scale studies, the identification of metabolites and lipids from this material will be important because no repeated analysis might be possible. Furthermore, standardized and reproducible methods with extensive coverage of metabolite and lipids are required.

The first question relates to the metabolome and lipidome of *C. elegans* and its constituents. So far, no *C. elegans* metabolome database exists. Therefore, it is crucial to compile metabolites that have been detected so far or are expected to present in the nematode. Chapter 2 covers this topic by reviewing different genome-scale metabolic models available for the nematode as well as their consolidation into a single consensus model. Several

manual curation steps were undertaken to add as much knowledge on *C. elegans'* metabolic pathways as possible. The end of chapter 2 summarizes a draft *C. elegans* metabolome and lipidome database, which will be used throughout the later chapters as the basis for metabolite and lipid identification. The reconstructed metabolic pathways are not only a blueprint to define the metabolites but also enable the interpretation of metabolomics and lipidomics results in the context of metabolic pathways. Chapter 3 describes approaches for the annotation and identification of metabolites and lipids in *C. elegans* with the aim of generating a reproducible workflow. This workflow is used in all following chapters for the annotation of the obtained metabolomics and lipidomics data. Such a workflow allows robust annotation and identification in large-scale experiments. The fourth chapter covers the identification of lipids in different obtained *C. elegans* lipidomics datasets. Although on the rise, lipidomics applications in *C. elegans* are still a minor part of *C. elegans* research, and even less is known about the lipidome composition. Lipids curated from previous lipidomics publications as well as lipids predicted from the WormJam GSMN, have been compared against lipids identified in three datasets. Taken together, they allow the construction of a draft lipidome database for *C. elegans*. This list of lipids enables fast dereplication in future *C. elegans* lipidomics experiments. Chapter 5 deals with the development of advanced analytical techniques, such as a tandem LC-MS setup for improved coverage in metabolite analysis. It combines RP and HILIC separation from a single injection and allows the analysis of polar and non-polar metabolites in a single analysis. Another method covers the analysis of fatty acids from single worms, which is an additional step towards single *C. elegans* metabolomics, allowing the study of the metabolic individuality of worms. The final chapter contains the conclusion of this work and gives an outlook on future steps required to establish *C. elegans* metabolomics and lipidomics in large-scale systems biology studies.

Several parts of the thesis have been published in different articles. Publications related to each chapter are indicated at the beginning of each chapter as well as my contribution to the individual publications.

- This page is left intentionally blank -

2. The *C. elegans* metabolic reconstruction and knowledge base

Chapter-related publications:

Modeling Meets Metabolomics-The WormJam Consensus Model as Basis for Metabolic Studies in the Model Organism Caenorhabditis elegans

Witting M., J. Hastings, N. Rodriguez, CJ. Joshi, JPN. Hattwell, PR. Ebert, M. van Weeghel, AW. Gao, MJO. Wakelam, RH. Houtkooper, A. Mains, N. Le Novère, S. Sadykoff, F. Schroeder, NE. Lewis, HJ. Schirra, C. Kaleta, O. Casanueva

Front Mol Biosci. 2018 Nov 14;5:96. doi: 10.3389/fmolb.2018.00096

This article describes the initial merging of several *C. elegans* genome-scale metabolic models into a consensus model called WormJam. With Prof. Dr. Horst Joachim Schirra, I'm leading the WormJam consortium and led the efforts for manual curation of metabolites, genetic information and reactions, especially regarding lipid metabolism. After the initial merge, I performed manual curation and correction of reactions, metabolites, and gene associations and added several *C. elegans*-specific pathways such as ascaroside biosynthesis.

Suggestions for Standardized Identifiers for Fatty Acyl Compounds in Genome Scale Metabolic Models and Their Application to the WormJam Caenorhabditis elegans Model

Witting M.

Metabolites. 2020 Mar 28;10(4):E130. doi: 10.3390/metabo10040130

This article describes a new nomenclature for acyl-based compounds, such as fatty acids, acyl-CoAs, acylcarnitines and others in genome-scale metabolic models. This nomenclature improves the readability of reactions and makes manual curation less tedious. I'm the sole author of this article, and I have developed the nomenclature and applied it to the WormJam consensus model.

2.1. Introduction

Knowledge about metabolites potentially present in an organism and their relation with each other, and their interconversion in metabolic pathways is crucial for correctly identifying metabolites and interpreting metabolomics results. The metabolism consists of a dense network of metabolic reactions, each catalyzed by one or several enzymes or enzymatic complexes. Several metabolic reactions form a pathway, and different pathways form a metabolic network. To study metabolism holistically, it is necessary to collect these reactions and the associated genes, enzymes, and metabolites for the specific organism that shall be studied. Still, many gaps exist, and often, metabolites not (yet) part of the metabolic network of the organism studied are identified by metabolomics and need to be integrated.

While genomics databases and protein databases are well developed and exist for different organisms or pan-organism, metabolism is underexplored [189]. Different types of knowledge bases or databases for metabolites exist. Metabolite and compound database store chemical information of small molecules and metabolites, such as names and synonyms, chemical structures, formulae, and other physicochemical properties. Metabolic pathway databases store information on reactions and enzymes catalyzing them. These databases are typically independent of organisms. Finally, Genome-Scale Metabolic Networks (GSMNs) are organism-specific depictions of metabolic pathways integrating information from the previously mentioned databases.

2.1.1. Metabolite and compound databases

Different metabolite databases exist and are used routinely in metabolomics. One of the most used databases is the Kyoto Encyclopedia of Genes and Genomes (KEGG) [190-193], which was also one of the earliest databases adopted for metabolomics. KEGG offers different interactive pathway maps and options to color them individually. They can be browsed, and each node is linked to the respective entry (e.g. a protein/enzyme or a metabolite). Different metabolomics annotation tools have used these pathways for the analysis of data [194-197]. However, for a few years, KEGG has a subscription model that

does not allow freely downloading information. KEGG contains information on *C. elegans* metabolic pathways and compounds, but the data must be retrieved manually, cannot be accessed programmatically, and is missing certain *C. elegans*-specific aspects.

The Human Metabolome Database (HMDB) stores information on metabolites from human metabolism and related substances, e.g. from the nutrition or the microbiome [86, 87, 198, 199]. At the current stage, this database contains 220,945 metabolites. These cover the complete range of polar to non-polar metabolites. Each metabolite is associated with rich metadata, such as formula, structure (e.g. SMILES, InChI, and InChIKey), and more. Furthermore, information on the concentration of metabolites in different biofluids is available for different conditions, e.g. healthy and diseased individuals. Therefore, HMDB also serves as reference value database. There is no direct link between human metabolism and metabolites stored in HMDB for *C. elegans*. However, it can be used to compare the metabolites in both organisms. The Yeast Metabolome Database (YMDB) follows the same structure as the HMDB and contains information on metabolites found in *S. cerevisiae* [200]. Likewise, the *E. coli* Metabolome Database (ECMDB) has been built around the *E. coli* K-12 strain [201]. The last one is of general interest for *C. elegans* since *E. coli* represent the standard diet in the laboratory for the worm, and metabolites produced by *E. coli* will also be ingested and are, therefore part of the *C. elegans* metabolome.

ChEBI is one of the best-curated metabolite databases of general interest and is independent of organisms [202, 203]. While other databases typically only store neutral, non-charged metabolite structures, ChEBI also differentiates between non-charged and charged versions of metabolites. This is particularly useful when working with metabolomics and GSMNs. While GSMNs contain molecular formulae and structures of metabolites present at the cytosolic pH of 7.3, in metabolomics experiments, neutral species are reported. Although scientists might refer to the same entity, in a chemical sense, they are not strictly the same, e.g. acetate is not acetic acid. Having different identifiers for the different charge species allows unequivocal identification of them. ChEBI links them together by a rich ontology, e.g.

acetate is the conjugate base of acetic acid. This ontology not only contains information on charge states but also on chemical classes and their relations. Furthermore, species-specificity is also stored in this ontology. Recently an approach was presented to use this ontology to map lipids to nodes in GSMNs [204]. The LipidMaps database stores information on different lipid classes and species based on a unique classification system. Lipids are classified into eight different classes: fatty Acyls [FA], glycerolipids [GL], glycerophospholipids [GP], sphingolipids [SP], sterol lipids [ST], prenol lipids [PR], saccharolipids [SL] and polyketides [PK] [205]. The LipidMaps structure database (LMSD) stores in total 47718 structures, from which 22043 are computationally generated and 25675 are curated from the literature. LipidMaps contains several lipids of interest for *C. elegans* and can be used as a database for annotation in *C. elegans*. SwissLipids is another database storing information on lipids [206]. In contrast to LipidMaps, which focuses only on reported lipid structures, SwissLipids used knowledge of lipid biosynthesis to generate potential lipids *in silico*. SwissLipids is curated by experts and links with other databases such as UniProt, ChEBI, Rhea, and Gene Ontology (GO). Lipid classes and their biosynthetic routes are curated from the literature and are based on the shorthand notation by Liebisch *et al.* [207]. At the moment, 779577 lipid species are covered in this database [206]. SMID-DB is a database of secondary metabolites in *C. elegans* and related species, such as *C. briggsae* or *P. pacificus* [140]. The goal is to supply shorthand identifiers for the large biochemical spaces covered by *C. elegans* secondary metabolites, e.g. different ascarosides. Recently, information on retention times and MS^2 has been added, making SMID-DB also an exciting resource for metabolite identification.

2.1.2. Metabolic reaction and pathway databases and Genome Scale Metabolic Models

Analysis of metabolomics data not only requires information on the metabolites but also on their relationship with each other. This information is typically stored in metabolic reaction and pathway databases. KEGG is not only a metabolite database but also contains information on metabolic reactions [190]. Genes and proteins are linked to enzymatic reactions. Similar to the metabolite database, the reaction database is not publicly available.

Reactions in KEGG are ordered in a hierarchical system loosely following the enzyme commission (EC) nomenclature, and association with species-specific genes and proteins are made. EC numbers, their hierarchy, and catalyzed reactions are also stored in the BRENDA [208]. RheaDB emerged as a recent alternative to storing information on metabolic reactions [209]. Mass and charge balanced reactions typically represent structures at the cytosolic pH of 7.3. Metabolite entries are thoroughly linked to ChEBI and catalyzing proteins to UniProt [189]. The Wikipathways database is a fast-growing collection of all kinds of pathways, including metabolic, interaction, and signaling pathways [210]. Reactome is another metabolic pathway database which contains information on metabolic pathways from 16 different organisms [211]. *C. elegans*-related entries cover 4376 proteins which are assembled into 4500 complexes. They catalyze 3587 reactions in 1296 pathways. MetaCyc is also manually curated and contains 2766 different pathways from 3067 different organisms [212]. However, it does not include information on *C. elegans'* metabolic pathways. The underlying software Pathway Tools was used for the construction of ElegCyc [213]. PathBank is a collection of human-curated pathways for ten different model organisms, including *C. elegans* [214]. Pathways were assembled with the PathWhiz illustrator. In total PathBank contains 110234 machine-readable pathways. Of these, 4244 belong to C. elegans, all being metabolic pathways. Information can be downloaded in different formats, e.g. SBML or BioPax

However, the above-mentioned pathways are of general interest, contain generic, integrated information from multiple species, and do not go deeper for single organisms. In order to close this gap and to enable better analysis of metabolism in biomedical model organisms, different metabolic reconstructions have been constructed from the genome of many organisms, including human, mouse, yeast, and others [215-217]. While the first GSMN of the gram-negative bacteria *Haemophilus influenzae* was rather small, currently, a trend towards more complex, multicellular organisms and full-body metabolic reconstructions is visible [218-220]. GSMNs represent the current knowledge on the metabolism of a given organism or superorganism. They group metabolites, genes, enzymes, and reactions

together to form a metabolic network that can be used to analyze metabolic networks *in silico* or used for the analysis of omics data. Typically, they are derived from *in silico* predictions from annotated genomes using different methods, followed by manual curation to improve annotations and close potential gaps networks can be used for various purposes [221-223].. An important step is the accurate depiction of individual molecular players used in the reaction network. Crucial is the correct curation of metabolites. They have to be represented at the correct charge state and correct molecular formula to define mass- and charge-balanced reactions.

As advancements in omics technologies continue to reveal the complexity of biological systems, different ways to analyze metabolism have evolved. Genetic manipulation of metabolic genes allows the study of their influence on different phenotypic traits, while metabolomics enables the characterization of the metabolic status of a given organism or biological system. Both techniques can be performed with decent throughput, but each only captures a small part of what is a genome-wide multi-omic system. Typically, metabolomics captures only a part of the metabolism from GSMNs by methodological constraints or stability of metabolites GSMNs and an accompanying technique, flux balance analysis (FBA), which allows the computational prediction of intracellular turnover rates (or fluxes) for all metabolic reactions in a cell or an organism, therefore overcoming some of the limitations of traditional techniques [224]. In addition, FBA can be informed by metabolomics results [52]. While GSMNs represent large knowledge bases of metabolism in a given organism, the discrepancies between these models and metabolomic measurements can be significant [20]. One of the key challenges for the use of GSMNs in conjunction with metabolomic data is the accurate mapping of measured metabolites to the metabolites present in the model. For example, metabolites found in a GSMN may fall below the limit of detection of current metabolomics techniques, or it may be present under specific conditions. Conversely, due to a large gap in the knowledge on metabolites and the large number of unknowns, many of them might be not yet be annotated and incorporated into the GSMN.

Different databases store GSMNs, e.g. BiGG contains a collection of GSMNs from various organisms [225]. MetExplore is a tool for the storage and collaborative curation of metabolic networks [226]. It hosts several metabolic networks from different organisms. Besides the actual storage, it offers rich visualization and allows mapping of different types of omics data to the network.

2.1.3. C. elegans metabolic reconstruction

For *C. elegans,* five different reconstructions have been published so far. The first one was published as part of the Path2Models approach to converting KEGG metabolic pathways into SBML models [227]. Models were gap-filled and checked if they were able to carry fluxes. In 2016 two metabolic reconstructions specifically focusing on *C. elegans* were published. Gebauer *et al.* performed reconstruction of the *C. elegans* metabolism using the PathwayTools software [213]. The final model consisted of 1921 reactions and 2357 metabolic compounds. The second model was published by Yilmaz *et al.* and used an own developed curation pipeline called SACURE [222]. The final model contained 1985 metabolic reactions and 1718 metabolic compounds. CeCon was published in 2017 and is also based on the PathwayTools software [228]. More recently, Worm1 as part of the reconstruction of GSMNs from multiple model organisms [229]. This is the by far currently the biggest model for *C. elegans* and has been derived from a human GSMN. Table 4 compares the different available *C. elegans* GSMNs. Since all these models are based on the same genome and the same genetic annotation, a community effort named WormJam was established which aims to reconcile and combine the different metabolic models into a single consensus reconstruction [230].

Table 4: Summary of published C. elegans GSMNs. Different GSMNs of different quality have been published so far. BMID000000141468 for example, was only based on automatic reconstruction, and no additional curation has been performed. Current Worm1 represents the largest model of C. elegans metabolism

Name	Compounds	Reactions	Reference	Notes
BMID000000141468	3207	2272	Büchel *et al.*, 2013 [227]	Automatic reconstruction from KEGG PATHWAYS, MetaCyc and gap-filling
ElegCyc axenic	2357	1893	Gebauer *et al.*, 2016 [213]	Bacteria-free growth media
ElegCyc E coli	2357	1921	Gebauer *et al.*, 2016 [213]	E. coli OP50, E.coli biomass composition (Orth et al., 2011) serves as the growth media.
iCEL1273	1718	1985	Yilmaz & Walhout 2016 [222]	
CeCon	2166	2085	Ma *et al.*, 2017 [228]	
Worm1	8150	12817	Wang *et al.*, 2021 [229]	Derived from Human GEM

2.2. Creation of The WormJam Consensus Metabolic Reconstruction

The reconstruction and maintenance of several models in inefficient and split the community effort towards a standardized representation of the current knowledge on *C. elegans* metabolism. For this reason, the WormJam community was built to enable the merging and further joint development of existing *C. elegans* metabolic models. Four of the published models have been integrated into the WormJam model at the current stage. These models have been merged automatically to form a version of a consensus reconstruction [231]. Several errors have been corrected, and new pathways have been added. The merging process is depicted in Figure 3.

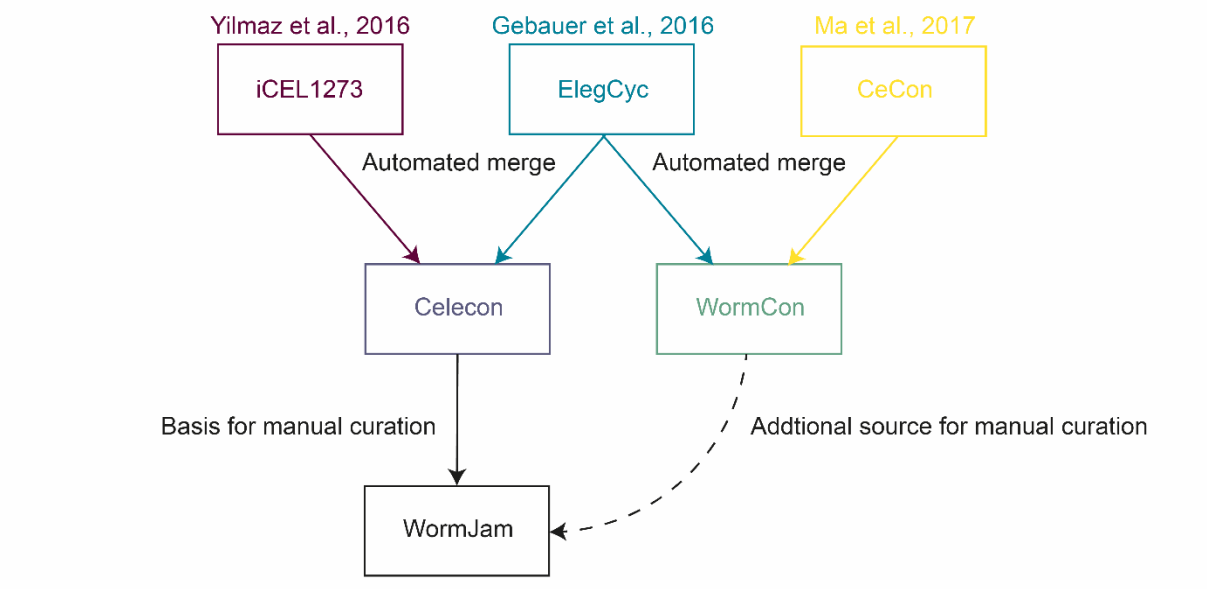


Figure 3: Depiction of the merging process to obtain the WormJam consensus metabolic reconstruction. iCEL1273 and ElegCyc merging has been performed by the group of Nathan Lewis and served the basis WormJam. The merged model WormCon derived from ElegCyc and CeCon served as additional information basis for manual curation.

On the basis of this model, several rounds of manual curation and correction of pathways have been conducted, which are partially presented in later paragraphs. A major step was the harmonization of names, identifiers etc., for genes, metabolites, and reactions since they were derived from different names spaces used by the various tools. Furthermore, new reactions and pathways have been added, and the curation of chemical structures was performed.

2.2.1. Reconciling Existing Models

In order to avoid merging too many models at once, a stepwise approach was used. Models based on the same curation tools or previously compared were merged first. By this already, several duplicated reactions and entries could be removed, leading to cleaner intermediate models for the final merging process.

2.2.1.1. Merging iCEL1273 and ElegCyc

The merging of these two previously published worm reconstructions ElegCyc and iCEL1273, involved using databases and standards for nomenclature since the two models used different identifier namespaces [213, 222]. For genes, first, a list of unique genes in each model was identified and linked. In WormJam, the WormBase gene identifiers (e.g. WBGene00001397 for *fat-5*) were selected for genes for the sake of their simplicity in parsing and formatting and accompanying ease of access to the online WormBase database [32]. If the gene was not found in gene mapping obtained from the GO website, KEGG was queried to obtain potential gene information [190]. The gene-protein-reaction association matrix was updated accordingly in case of duplicated gene entries due to the different namespaces. Gene rules are encoded in a text field describing the relation of genes with each other. Isoenzymes are separated by an “or” while subparts of multimers are separated by an “and”.

Similar to genes, several different namespaces for metabolites were used. Unique metabolites were identified by using BiGG, KEGG, and MetaNetX [190, 232, 233]. For metabolites present in BiGG, additional information such as ChEBI, KEGG, and MetaNetX IDs were extracted together with chemical formulae and charge states. In cases where no information was found, a manual search was performed. Duplicate metabolite entries within the model were fixed by removing one of the instances and resolving in-model metainformation and stoichiometry.

An automatic script was used to perform a comparison of reactions (available from https://github.com/LewisLabUCSD/celegans_reconciliation). If reactions were found to be the

same, the reaction from the first model was kept and updated with the information from the second. Furthermore, stoichiometry was compared and checked. This initial list was further used for manual curation steps explained below.

2.2.1.2. Merging of CeCon and ElegCyc

The merging of CeCon and ElegCyc has been performed using the COMMGEN toolbox [234]. Both models were translated to the MetaNetX namespace to perform semi-automated merging [233]. After completely merging the two models into a single COMMGEN entity in MATLAB, the biomass reaction of ElegCyc was set as the objective function for flux balance analysis testing of the model's validity, as CeCon, which was previously only used for the mapping of "-omics" data, does not contain a biomass reaction. Next, the following steps were performed semi-automatically through COMMGEN functions.

```
1. Merging of duplicate reactions
2. Merging of reactions with similar species
3. Removal of nested reactions
4. Alteration of invalid transport reactions
5. Removal of invalid external reactions
6. Checking of reactions with the same metabolites, but differing
stoichiometry
7. Merging of similar transport reactions
8. Merging of duplicate reactions
```

In each function, COMMGEN suggested potential merge candidates, and a decision was manually made on whether to merge the reactions or keep both based on literature and databases such as KEGG [190]. After every merge, the model was exported to COBRA Toolbox format and checked for viability by ensuring flux was able to be carried through the merged model's biomass reaction in a "free growth" simulation. If the merges were inviable, the change would be reverted. At the end of this process, the resulting model, named WormCon, was exported to SBML format, and further merging was halted in favor of manual curation as part of WormJam.

2.2.2. *WormJam Infrastructure*

After merging all available models into a single consensus model, future curation steps should be streamlined and harmonized to avoid scattering of efforts and duplication of curation steps. The WormJam consensus model is maintained on a dedicated GitHub page with public access (https://github.com/wormjam-consortium/wormjam). Although originally developed for version control and tracking in software development, this system can be used for virtually any text-based file format. This allows automated monitoring of all changes made to the model. The use of GitHub follows the recently published Yeast8 consensus model for *S. cerevisiae* [217]. Besides the actual model files, several utility scripts written in Python, Java, or R are stored with the model, which are used for data cleaning and quality control.

Metabolic models are typically stored in the SBML format, an XML derivative customized for use in systems biology [235-237]. However, this format is hardly human-readable and contains a lot of information to aid the computer in processing the file. In contrast to SBML files, the Systems Biology Table (SBtab) format allows humans to work directly with the model, which is of great advantage for manual curation. Both formats are interconvertible, and free web tools for conversion are available (https://www.sbtab.net/sbtab/default/converter.html) [238]. The SBtab format comprises a series of separate files that individually contain the reactions, metabolites, genes, and other information present in the GSMN. Because the data in the individual data files are related to each other, the same identifiers are used across all individual files. This makes it possible to cross-reference and search for information e.g. isolate metabolites from a single reaction and retrieve information on them from the compound. The WormJam data in the SBTab format is organized in different tables, summarized in Table 5.

In order to enable further development of the model without interfering with stable model versions, different branches are used in the GitHub repository. The `master` branch contains the major release versions, while the `devel` branch allows the generation of new model versions for curation and testing. Different users have to create their own branches and

contribute via the git system using commits containing their changes. Once finished, pull requests into the `devel` branch are performed. These pull requests are checked automatically and reviewed by experts to avoid inconsistencies between branches. A pull request triggers a series of GitHub actions, which perform the conversion of the SBtab files into SBML, followed by automatic checks using the `memote` test suite for GSMNs [239]. Most importantly, the charge and mass balance of reactions and duplicated reactions are reported. Results are then reported to the user/curator for further checking. If the model is consistent, a PR into the main repository can be performed, and changes will be integrated into the `devel` branch of the main model. Finally, after the model is sufficiently advanced, a development version can be merged into the `master` branch leading to a new release version. Since all changes are logged, potential errors can be tracked for corrections.

Table 5: Different tables of the WormJam model in SBtab format. The tabular and human-readable format allows easy manual curation. Different parts of the model are stored and linked via unique identifiers for reactions, genes, metabolites, curators, etc.

Table	Content
Compartment-SBtab.tsv	Contains definitions of compartments in the model
Compound-SBtab.tsv	Contains all information on metabolites, incl. charge state, formula, chemical structure
Curator-SBtab.tsv	Contains information on involved curators and their eMail addresses
Definition-SBtab.tsv	Contains information on the individual columns present in all other SBtab files
Gene-SBtab.tsv	Contains information on all involved genes, including gene symbols, WormBase gene identifier etc
Pathway-SBtab.tsv	Contains information on pathways and how they relate to pathways from other databases
Reaction-SBtab.tsv	Contains all metabolic reactions

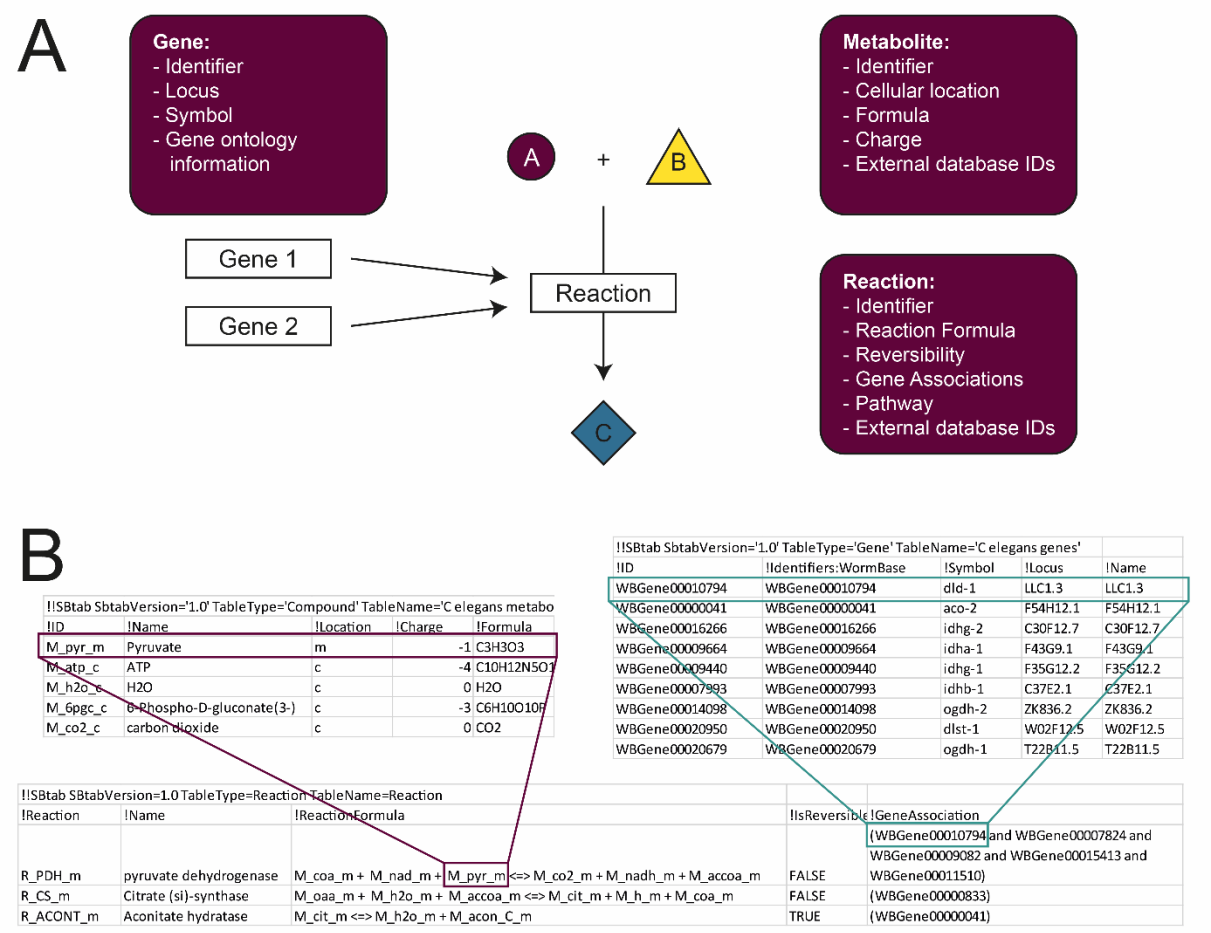


Figure 4: (A) The three most common classes of entities in GSMNs are genes, metabolites and reactions. Gene association to reactions forms the basis of GSMNs and metabolites can be either substrates or products in reactions In this representation, Genes 1 and 2 are metabolic genes required for the catalysis of the reaction, which converts A and B into C. (B) The used SBTab format is based on distinct tables for each entity (compare Table 5). Example of indiviual tables from SBTab format linked entries between tables are highlighted.

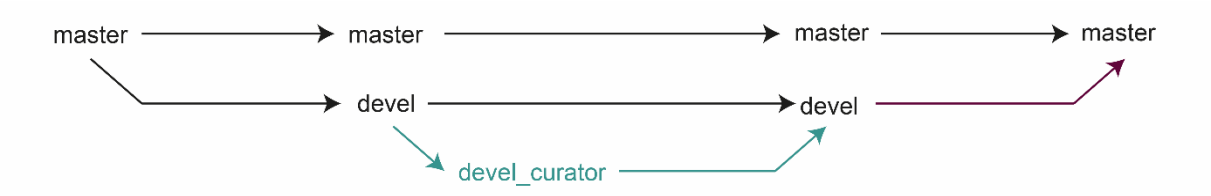


Figure 5: GitHub is used for the storage of different versions of the model. From a master branch (blessed repository) different developmental version can be generated, which allow the public or private development in the model. By pull requests, new changes are integrated once tested. The master branch is only changed by integration managers to avoid introduction of errors or duplications.

2.2.3. Metabolite structure curation, namespace, and metabolite nomenclature

Standardizing the representation of metabolites in GSMNs is an important issue to enhance reuse (also for scientists not involved in developing the specific model) and facilitate comparison across different GSMNs. Furthermore, using the same and consistent name space allows merging other models into larger models (e.g. host-microbiome models).

Several GSMNs have been integrated into the WormJam consensus model, each with a different identifier namespace, metabolite naming, and detail of structural curation [213, 222, 228, 231]. Therefore, after automatic merging, conflicts in namespaces and potential duplications had to be resolved. In GSMNs, often only the chemical formula is curated since the exact chemical structure is not essential for typical modeling issues, such as FBA. Furthermore, several new metabolites have been introduced during the manual curation steps, which must be integrated into existing namespaces (see 2.2.5 and following). Lastly, integration of GSMNs specific databases such as BiGG or MetaNetX is important for cross-model comparison [232, 240]. While each metabolite has a unique chemical structure, which allows its unambiguous identification, GSMNs often rely on simple identifiers, and sum formulae and only a few metabolic reconstructions have detailed curation of chemical structures associated with the model. Examining nearly 100 published GSM, Ravikrishnan and Raman found that over 60% of these models were lacking standard metabolite identifiers such as KEGG IDs, PubChem IDs, or InChIs [241]. Currently, a significant community effort toward the standardization of metabolic models has been started [242, 243].

Detailed structural curation is performed in WormJam. GSMNs use the major microspecies of metabolites at the cytosolic pH of 7.3 metabolite structures and mass and charged balanced reactions, i.e., reactions in which atoms are neither created nor destroyed. One example is acetic acid, which is present as acetate in the model. In a chemical sense, the two are separate entities with different chemical properties. The ChEBI database also stores the two in individual entries, while KEGG, for example, contains only one single entry for acetic acid and acetate stored as a synonym. Therefore, WormJam uses ChEBI as the central repository for metabolite structures [202].

Both, charged and uncharged, structures are curated within the WormJam model and have been added to ChEBI if they were missing. ChEBI IDs served as reference identifiers for the charged and neutral structures. Still, KEGG, MetaCyc, HMDB, LipidMaps, SwissLipids, Wikidata, PubChem, Metabolights, and Chemspiders IDs have been additionally curated for

the neutral metabolite version to allow cross-mapping between databases. IDs are fetched from the most current version of BridgeDB using the BridgeDbR package [244, 245]. Searches in BridgeDB are based on the InChIKey of the neutral molecule. To not have to use the full chemical name in reactions, metabolite identifiers are used, and WormJam uses identifiers from BiGG [232]. If no identifiers have been found in BiGG, new ones have been created. According to the BiGG models ID specification and guidelines, IDs for metabolites should be human-readable, short, and memorable. However, when curating reactions manually, it is also essential to derive (at least partial) structural information from the ID to simplify the manual curation process. This is especially true for long reaction sequences, e.g. biosynthesis or β-oxidation of fatty acids. Inconsistencies between different namespaces from parent models have mainly been observed for fatty acids, acyl-CoAs, acyl-Carnitines, and acyl-ACPs. Although the different classes often had consistent naming within one class, it was not always directly comparable to other classes, e.g. acyl-CoAs and acyl-carnitines. Searching for reactions and metabolites, it was found that different classes of metabolites required for the biosynthesis and elongation of fatty acids and β-oxidation had very different "namespaces" and used various abbreviations for the same acyl moiety. Inspecting IDs used in other BiGG models, different levels of structural details in the different IDs and naming have been observed. IDs and chemical names of acyl-based metabolites have been downloaded from BiGG, and where available, associated chemical information, such as ChEBI IDs etc., was used. Interestingly, many metabolites had no structural information associated with the BiGG DB nor IDs from chemical structure databases.

First, an overview of the data available from BiGG was generated. In order to evaluate if IDs and naming enable unambiguous identification of the molecular entities within the different classes, they were grouped into different categories. The first group was called "full structural information" if structural information could be derived from the ID, name, or links to unique chemical structures were available, "partial structural information" if minor information was missing, or "no structural information" if major information, e.g. position and/or stereochemistry of a double bond or functional group was missing. An example of an ID and

name pair that was classified as “full structural information” is *lnlccoa*, which is the ID for Linoleyl-CoA. *ttdcrn*, tetradecanoyl carnitine, was classified as “partial structural information” since the stereochemistry of the carnitine is missing. In contrast, *ttdcea*, Tetradecanonate (n-C14:1) is classified as “no structural information” because important information like the position and stereochemistry of the double bond was absent.

For fatty acids, acyl-CoAs and ACPs from BiGG >66% had complete structural information available. Acyl-carnitines showed the highest number of partial structural details, mostly because the stereochemistry of the carnitine moiety was missing. Analysis of names, IDs, and potentially linked chemical information revealed that for most metabolites, structural information could be deduced. Still, several metabolites with only partial or no structural information are present. To explore if annotations of metabolites grouped as “no structural information” can be improved, the reactions in which these metabolites occur were checked. It might be possible that an upstream metabolite contained the full structural information, but this information was not propagated correctly. As an example, the metabolite *ttdcea*, annotated as tetradecenoate, was searched for in BiGG. This metabolite is used in many different models. The *E. coli* model iJO1366 was selected, where this metabolite occurs in the cytosol and is linked to 9 reactions [23]. Selecting the reaction “FA141ACPHi” (Fatty-acyl-ACP hydrolase), it could be seen that *ttdcea* is derived from a more detailed *tdeACP*, which is described as “cis-tetradec-7-enoyl-[acyl-carrier-protein] (n-C14:1)”. Using this information *ttdcea* should be annotated as “cis-7-tetradecenoate acid” instead of “Tetradecenoate (n-C14:1)”. “Cis-7-tetradecenoate” can be found as metabolite M00117 in BiGG derived from Recon 3D [246].

Based on the idea that the reaction network can be used to check for connections between identifiers of different quality, reactions from BiGG were used to isolate pairs of acyl-based metabolites as they occur as substrates and products of reactions. All pairs that contained hub metabolites such as CoA, Acetyl-CoA, Carnitine, or ACP and pairs derived from transport reactions were removed. Metabolites were then labeled according to their group,

and the connection of metabolites of different groups was checked. Following counts for pairings of the different categories were found: no structural information<-> no structural information = 135, partial structural information <-> no structural information: 8, partial structural information <-> partial structural information: 36, full structural information <-> no structural information: 96, full structural information <-> partial structural information: 161, full structural information <-> full structural information: 897. Based on these pairings, it was determined that metabolites grouped into the category “full structural information” are mostly connected to metabolites of the same group. However, also several connections between metabolites classified as “full structural information” and “no structural information” metabolites exist. These connections with known biochemistry can be used to improve the annotation of metabolites with no structural. Furthermore, plotting all pairings as a network showed several long-distance possibilities for improvement exists. As a result of this comparison the annotation of specific metabolites can be improved. A particular example is the metabolite *arachdcoa_c*, which is annotated as “C20:4-CoA”. However, the identifier suggests that this metabolite might be “Arachidonoyl-CoA”; no supporting information is supplied. Looking into connected metabolites, *arachd_c* was identified as a metabolite classified as “full structural information”. The name “Arachidonic acid” identifies it with specific positions and stereochemistry of the double bonds (FA 20:4(5Z,8Z,11Z,14Z)), which identifies *arachdcoa_c* as “arachidonyl-CoA”. As a second reaction in links *arachdcoa_c* to *adrncoa_c*, “adrenyl-CoA” is also grouped as “full structural information”, which in turn is linked again to *adrn_c*, “adrenic acid”. This example shows how the network can be used to improve the structural annotation of different metabolites to full structural detail.

Results have shown a considerable heterogeneity in the use of IDs with different degrees of structural curation. Although several IDs exist for which complete structural information can be retrieved, even IDs that have full structural details from other models are very different in the way the encode this structural information. All investigations have shown that a more systematic way of naming IDs of such acyl-based metabolites is required to avoid future confusion.

In lipidomics, the use of shorthand notations instead of long, systematic IUPAC names is widespread to describe lipids. This shorthand notation has been standardized by Liebisch *et al.* and is the most widespread notation, continuously improving to cover more lipid classes and modifications [207, 247]. Such a systematic notation would resolve problems of cross-mapping between models; however, it is not compliant with the specifications for BiGG metabolite identifiers. Transfer and adoption of this nomenclature for BiGG IDs will improve the readability of IDs as well as structural details. A similar idea was already followed in the Chinese hamster ovary cell consensus model iCHOv1, where, for example, c81_5Zcrn_m was used to describe a carnitine molecule with a C8 chain and a cis-double bond at position 5, but no rules how to generate the IDs nor an automatic way have been proposed [248].

A typical example of a shorthand notation used in lipidomics to describe a fatty acid or acyl-CoA is shown in Figure 6. Structural features of an acyl chain are encoded in human-readable abbreviations. First, the class of the molecule is denoted, and afterwards, structural features are encoded. Next, the number of carbons of the longest chain is followed by the number of double bonds separated by a ":". In a pair of brackets, structural details like the position and geometry of double bonds or the position and potential stereochemistry of functional groups are given. At the current stage, the following functional groups are supported in the lipidomics shorthand notation: keto (O), hydroxy (OH), peroxy (OOH), amino (NH2), and methyl (Me) groups. If the groups have stereocenters, it is defined directly after the functional group in square brackets, e.g. OH[S]. Each group additionally needs a number indicating the position, for example, 3OH[S]. Individual groups are separated by a comma. The order of function groups is double bonds, hydroxy groups, peroxy groups, keto groups, amino groups, and methyl groups (DB > OH > OOH > O > NH2 > Me).

Similar rules are also applied to generate IDs for GSMNs following the guidelines for valid BiGG IDs. First, the class of the molecule is denoted in lower case letters directly followed by the number of carbons and the double bonds separated by an underscore "_" (e.g. fa18_2). After a second underscore "_" the functional groups are followed as a combination of position

and lower-case letters with no separator (e.g. fa18_2_9z12z). If the functional group, e.g. a hydroxyl group, represents a stereocenter, the stereochemistry is directly indicated with the functional group as a capital letter (e.g. coa18_0_3ohS). Stereocenters of the base molecules, such as carnitine, are denoted similarly to amino acids at the end with two underscores "__" (e.g. carn18_1_9z__L). Nomenclatures for stereocenters shall not be mixed, and only either R/S or D/L nomenclature should be used within one identifier.

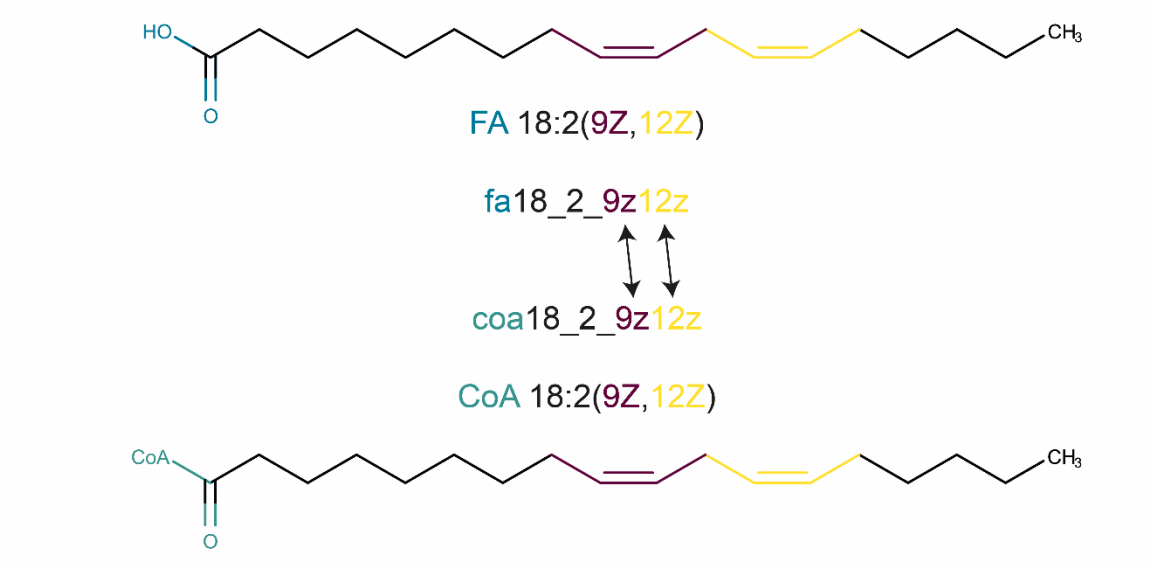


Figure 6: Example of structures, shorthand notation based on Liebisch et al. and the systematic nomenclature for acyl-based IDs.

The use of the newly suggested identifiers should follow certain rules. First, the abbreviation shall be only used for metabolites with more than five carbons (>= six carbons), which avoids the need to change IDs for metabolites involved in pathways other than fatty acid metabolism, e.g. central carbon or amino acid metabolisms such as acetyl-CoA (*accoa*), isovaleryl-CoA (*ivcoa*) or others. Second, if an abbreviation of a trivial metabolite name is easier to understand and allows unambiguous identification, this abbreviation should be used. For example the oxidized fatty acid 5(S),15(R)-DiHETE (CHEBI:91138, FA(20:4(6E,8Z,11Z,13E,5OH[S],15OH[S])) shall be abbreviated as dihete_515__SR instead of fa20_4_6e8z11z13e5oh15oh__SR. BiGG contains only a generic metabolite ID for this metabolite CE7096, which shall be renamed to a generic version dihete_515. Third, it is suggested to use this nomenclature, especially in long reaction sequences like fatty acid

biosynthesis, elongation, or β-oxidation. Fourth, consistent naming shall be used across the model for all classes of acyl-based metabolites, e.g. fatty acids should be named the same way as acyl-CoAs, e.g. fa18_1_9z should be used together with coa18_1_9z and not with coa18_1. If generic metabolites like coa18_1 shall be used in the model, pseudo reactions converting explicit metabolites to generic versions (coa18_1 <=> coa18_1_9z) have to be added. An example of the different identifiers before and after harmonization is given in Table 6.

Table 6: Comparison of different identifier for acyl-based metabolomics from iCEL1273 and ElegCyc as parent model of WormJam as well as from the WormJam model before the manual curation step and BiGG. The identifiers have been transferred to the new more systematic identifier.

Metabolite	iCel1273	ElegCyc	WormJam	BiGG	New id
Tetradecanoic acid (Myrstic acid)	ttdca	ttdca	ttdca	ttdca	fa14_0
Tetradecanoyl-Acyl-Carrier Protein (Myristoyl-ACP)	Myristoyl_ACPs	---	myrsACP	myrsACP	acp14_0
Tetradecanoyl-Coenzyme A (Myristoyl-CoA)	tdcoa	tdcoa	tdcoa	tdcoa	coa14_0
Tetradecanoyl-Carnitine (Myristoyl-Carnitine)	ttdcrn	CPD909_16	ttdcrn	Ttdcrn M02973	carn14_0
Tetradecanoyl-Ethanolamide (Myristoyl-ethanolamide)	---	---	---	tetdecaeth	nae14_0

Similar to acyl-based metabolites, identifiers of lipid classes suffer from limited systematic and comparability. Although several identifiers from BiGG were reused wherever possible, new identifiers were required to be introduced in some instances. This includes IDs for complex lipids, such as PCs, PEs, etc. Since BiGG shows some inconsistencies in the naming of lipids, the nomenclature of different lipid classes within WormJam was revised, and clearer IDs have been introduced, which enhance readability and clearness of reactions

and allow more effortless manual curation. For example, after merging, the WormJam model contained different lysolipids with no further defined structure (e.g. sn1 and sn2 lysolipids). The generic metabolites M_acg3pe_c represented both 1-Acyl-sn- or 2-Acyl-sn-glycerol-3-phosphoethanolamine. Therefore, no specificity could be encoded in the reactions. An additional example is 1-acyl-sn-glycerol-3-phosphate for which 3 IDs are used dependent on the biosynthetic origin. The intention behind this was to generate different pools from which PA can be derived, but as understood so far, cells and organisms generally do not differentiate, and all molecules are part of the same pool. The new nomenclature of the lipids IDs is based on different sn-glycero-phospho-lipids, like sn-3-glycero phosphocholine or sn-3-glycero-phosphoethanolamine, which use standard abbreviations such as g3pc or g3pe. In the new nomenclature, acyl, alkyl, or alkenyl moieties are abbreviated as ac, alk, or alken, respectively. A number indicates the position. Therefore, the metabolite 1,2-diacyl-sn-glycero-3-phosphocholine would be abbreviated as 1ac2acg3pc. All new IDs can be found in Table 7. To allow back tracing and backward compatibility of the new metabolites IDs, in the compound table, a new column called "!Notes:Old_ID" has been added, which holds previously used identifiers. Based on the suggested new nomenclature IDs within WormJam have been updated, and duplicated entries in both metabolites and reactions have been identified and removed.

Table 7: Table of corrected or newly introduced IDs for different lipid classes. In GSMNs only generic nodes for entire lipid classes exists. Different models used in WormJam used different nomenclature. This has been harmonized to a common systematic nomenclature.

Class	Metabolite	Old/ Wrong / Duplicated ID	Correct / New ID
MG	1-acyl-sn-glycerol	1magol	1acglyc
	2-acyl-sn-glycerol	mag	2acglyc
MG-O	1-alkyl-sn-glycerol	---	1alkglyc
MG-P	1-(Z)-alk-1-enyl-sn-glycerol	alkenglyc	1alkenglyc
DG	1,2-diacyl-sn-glycerol	12dag	1ac2acglyc

Class	Metabolite	Old/ Wrong / Duplicated ID	Correct / New ID
DG-O	1-alkyl-2-acyl-sn-glycerol	akac2g	1alk2acglyc
DG-P	1-(Z)-alk-1-enyl-2-acyl-glycerol	alkenac2g	1alken2acglyc
TG	Triacyl-glycerol	tag	tag / 1ac2ac3acglyc
TG-O	1-alkyl-2,3-diacylglycerol	---	1alk2ac3acglyc
TG-P	1-(Z)-alk-1-enyl-2,3-diacylglycerol	---	1alken2ac3glyc
DHAP	1-acylglycerone 3-phosphate	Adhap	1acdhap
DHAP-O	1-alkylglycerone 3-phosphate	akdhap	1alkdhap
PA	1,2-diacyl-sn-glycero-3-phosphate	pa_pl 12dag3p	1ac2acglyc3p
LPA	1-acyl-sn-glycero-3-phosphate	alpa alpa_tag 1ag3p_SC	1acglyc3p
	2-acyl-sn-glycero-3-phosphate	---	2acglyc3p
LPA-O	1-alkyl-sn-glycero-3-phosphate	alkgp	1alkglyc3p
PA-O	1-alkyl-2-acyl-sn-glycero-3-phosphate	akac2gp	1alk2acglyc3p
LPA-P	1-(Z)-alk-1-enyl-sn-glycero-3-phosphate	---	1alkenglyc3p
PA-P	1-(Z)-alk-1-enyl-2-acyl-sn-glycero-3-phosphate	---	1alken2acglyc3p
PC	1,2-diacyl-sn-glycero-3-phosphocholine	pchol	1ac2acg3pc
LPC	1-acyl-sn-glycero-3-phosphocholine	ag3pc	1acg3pc
LPC	2-acyl-sn-glycero-3-phosphocholine	2agpc	2acg3pc
PC-O	1-alkyl-2-acyl-sn-glycero-3-phosphocholine	akac2gchol	1alk2acg3pc
LPC-O	1-alkyl-sn-glycero-3-phosphocholine	ak2lgchol	1alkg3pc
PC-P	1-(Z)-alk-1-enyl-2-acyl-sn-glycero-3-phosphocholine	---	1alken2acg3pc
LPC-P	1-(Z)-alk-1-enyl-sn-glycero-3-phosphocholine	---	1alkeng3pc
PE	1,2-diacyl-sn-glycero-3-	pe	1ac2acg3pe

Class	Metabolite	Old/ Wrong / Duplicated ID	Correct / New ID
	phosphoethanolamine	pe_BAC	
LPE	1-acyl-sn-glycero-3-phosphoethanolamine	acg3pe	1acg3pe
LPE	2-acyl-sn-glycero-3-phosphoethanolamine	---	2acg3pe
PE-O	1-alkyl-2-acyl-sn-glycero-3-phosphoethanolamine	akac2gpe	1alk2acg3pe
LPE-O	1-alkyl-sn-glycero-3-phosphoethanolamine	---	1alkg3pe
PE-P	1-(Z)-alk-1-enyl-2-acyl-sn-glycero-3-phosphoethanolamine	alkenac2gpe	1alken2acg3pe
LPE-P	1-(Z)-alk-1-enyl-sn-glycero-3-phosphoethanolamine	alken2gpe	1alkeng3pe
PS	1,2-diacyl-sn-glycero-3-phospho-L-serine	ps	1ac2acg3ps
LPS	1-acyl-sn-glycero-3-phospho-L-serine	acg3ps	1acg3ps
	2-acyl-sn-glycero-3-phospho-L-serine	---	2acg3ps
PI	1,2-diacyl-sn-glycero-3-phospho(1)-D-myo-inositol	pail	1ac2acg3pi
LPI	1-acyl-sn-glycero-3-phospho(1)-D-myo-inositol	---	1acg3pi
	2-acyl-sn-glycero-3-phospho(1)-D-myo-inositol	---	2acg3pi
PIP	1,2-diacyl-sn-glycero-3-phospho(1)-D-myo-inositol-3-phosphate	pail3p	1ac2acg3pi3p
	1,2-diacyl-sn-glycero-3-phospho(1)-D-myo-inositol-4-phosphate	pail4p	1ac2acg3pi4p
	1,2-diacyl-sn-glycero-3-phospho(1)-D-myo-inositol-5-phosphate	pail5p	1ac2acg3pi5p
PIP2	1,2-diacyl-sn-glycero-3-phospho(1)-D-myo-inositol-3,4-bisphosphate	pail34p	1ac2acg3pi3p4p
	1,2-diacyl-sn-glycero-3-phospho(1)-D-myo-inositol-3,5-bisphosphate	pail35p	1ac2acg3pi3p5p
	1,2-diacyl-sn-glycero-3-phospho(1)-D-myo-inositol-4,5-bisphosphate	pail45p	1ac2acg3pi4p5p
PIPI3	1,2-diacyl-sn-glycero-3-phospho(1)-D-myo-inositol-3,4,5-trisphosphate	pail345p	1ac2acg3pi3p4p5p
PGP	1,2-diacyl-sn-glycero-3-phospho-(1'-sn-glycero-3'-phosphate)	pgp	1ac2acg3pg3p

Class	Metabolite	Old/ Wrong / Duplicated ID	Correct / New ID
PG	1,2-diacyl-sn-glycero-3-phospho-(1'-sn-glycerol)	pg pg_BAC	1ac2acg3pg

2.2.4. Charge states

Metabolites in GSMNs are used as charged versions, and reactions are typically mass and charged-balanced. Charges state a pH of 7.3 is calculated using the major microspecies present at this pH. Incorrect charge states influence the reaction balance, leading to false results in FBA. In order to validate the charge states of all metabolites, an Excel sheet, which utilizes functions from JChem for Excel, was developed. Chemical structures based on the SMILES representation are used as input. In a specific part, the complete compound table in SBtab format is pasted, and all values are calculated automatically and compared against the charge states currently present in WormJam.

First, based on the given neutral SMILES for a molecule, the major microspecies at pH 7.3 is calculated. Afterward, for this microspecies, the charge state is calculated. In parallel, the charge state is calculated from the SMILES of the charged species given in the compound table. If the two calculated charge states and the charge state in the compound table match, no action is required; otherwise, charge states need to be verified manually and corrected. After structural curation of the different metabolites, this sheet is used to validate correct charge states or calculate charge states for new molecules. For all metabolites, the charge state has been recalculated, and the correct charge state as well as the correct charged structure (major microspecies), added to the table. In parallel charge states have been cross-validated using BiGG. In case of inconsistency, entire reactions were checked to use the most consistent charge states.

2.2.5. Metabolic pathway curation

After the initial automatic draft reconstructions and the automatic merging of the different models into the WormJam model, additional steps of manual curation have been performed. This included the curation of missing metabolic pathways and reactions, correction of wrong

gene assignments, and others. Extensive literature is available on different metabolic pathways in *C. elegans*. However, often this knowledge is not reflected in electronic databases or automated reconstructions. Therefore, the manual curation of metabolic pathways represents an important step towards better coverage of the *C. elegans* GSMN and improves the understanding of metabolism in this model organism.

Repetitive reaction sequences, such as fatty acid elongation or β-oxidation are often shortened into lumped reactions. This is a convenient way of reducing the number of reactions but does not reflect the true reaction details. Often several genes are then associated with one lumped reaction, which does not directly allow assessment of effects, e.g. *in silico* knock-out simulations. A major goal was the removal of lumped reactions since the use of the AND notation for different genes shall be used only for complexes. This operator shall be only used to account for reactions that require multiple gene products, e.g. multienzyme complexes, etc. The switch to detailed reactions instead of lumped reactions also allows better to identify effects of gene mutations or flux analysis. The next paragraphs discuss the curation of different pathways which needed improvement or have been completely missing. This includes the correction of already existing reactions and pathways or the addition of novel ones.

2.2.5.1. Propionate metabolism

Vitamin B12 is an important molecule and cofactor in many enzymes, including enzymes required for the degradation of the short-chain fatty acid propionate. Under B12 deficiency, an alternative pathway is used, which produces Acrylyl-CoA, which in turn can hydrolyze to produce toxic acrylate [249]. Most of the reactions were present in the WormJam model and had correct assignments but were missing the assignment to the right pathway definition. Only the hydroxy propionate dehydrogenase reaction was missing the correct gene assignment.

2.2.5.2. Fatty acid biosynthesis

Fatty acids represent important building blocks in, e.g. different lipids or secondary metabolites such as ascarosides. *C. elegans* takes up fatty acids from its food and incorporates nutritional fatty acids into its biomass. However, different fatty acids can be synthesized by the worm itself *de novo*, including BCFAs. These fatty acids are almost exclusively (>99%) produced by the worm itself [123]. BCFAs are important molecules and are required for the biosynthesis of sphingoid bases in *C. elegans* [116, 250, 251]. The biosynthesis of fatty acids in *C. elegans* can be split into different steps. First, palmitic acid (FA 16:0) or 11-methyl dodecanoic acid (FA 12:0(11Me)) are produced by FASN-1 in 7 or 5 cycles of 4 fatty acid synthesis, respectively. Both are then further elongated and desaturated to produce different other fatty acids.

The merged model initially showed different details of the curation of fatty acid metabolism. In the frame of the WormJam reconciliation, the level of detail was harmonized, and all intermediate steps were added. Fatty acid biosynthesis is well established based on the work of Watts and Browse [252]. In their seminal paper, Watts and Browse constructed a biosynthetic pathway for PUFAs in *C. elegans*. Using GC-FID in combination with an EMS screen, they identified several genes that alter the fatty acid composition [252]. As the first step, all fatty acid-related metabolites have been curated to the new nomenclature for acyl-based metabolites as described above. This harmonized the namespace for all fatty acyls and allowed the identification of duplicated reactions for downstream curation. Fatty acid biosynthesis covers different aspects, e.g., synthesis of straight chain even numbered fatty acids as well as iso-branched chain odd numbered fatty acids. Searching for reactions associated with branched chain fatty several lumped/nested reactions have been identified, e.g. the following reaction which produces the fatty acid 11-methyldodecanoic acid (FA 12:0(11Me)) catalyzed by FASN-1 (WBGene00009342). The following reaction formula was found:

```
11.0 M_h_c + 8.0 M_nadph_c + M_ivcoa_c + 4.0 M_malcoa_c
<=>
```

4.0 M_co2_c + 3.0 M_h2o_c + 5.0 M_coa_c + 8.0 M_nadp_c + M_fa13p0iso_c

This reaction is stoichiometrically correct but misses several important aspects of fatty acid biosynthesis, e.g., the Acyl-Carrier-Protein (ACP) involvement. Biosynthesis is performed within a complex, and the growing fatty acyl is covalently bound to ACP. Checking for further elongation following reactions were found, elongation FA(12:0(11Me)) to FA(14:0(13Me)) and FA(16:0(15Me)).

3.0 M_h_c + 2.0 M_nadph_c + 2.0 M_malcoa_c + M_fa13p0iso_c
<=>
M_co2_c + M_h2o_c + M_coa_c + 2.0 M_nadp_c + M_fa15p0iso_c

3.0 M_h_c + 2.0 M_nadph_c + M_malcoa_c + M_fa15p0iso_c
<=>
M_co2_c + M_h2o_c + M_coa_c + 2.0 M_nadp_c + M_fa17p0iso_c

Likewise, these reactions miss the critical fact that elongation reactions are carried on Acyl-CoAs, not free fatty acids. All these reactions were inferred from iCel1273 because ElegCyc did not contain any reactions related to BCFA biosynthesis. To correct these errors and have a more detailed depiction, reactions related to BCFAs were created from scratch using textbook knowledge and published gene associations. All previous (lumped) reactions have been deleted. In parallel, the biosynthesis and elongation of straight chain fatty acids were also corrected. These reactions were also missing the involvement of ACP. Reactions have been separated into individual steps also to have unique gene-reaction relationships. Furthermore, *C. elegans* also contains mitochondrial fatty acid biosynthesis. To allow more specific pathway assignment, reactions are sorted into pathways depicted in Table 8: Number of reactions associated with newly generated metabolic pathways related to fatty acid biosynthesis, elongation and desaturation.

Reactions for each step of fatty acid biosynthesis have been added, including condensation, 1st reduction, dehydration, and 2nd reduction. The specific Acyl-ACPs have been added to the metabolite list. Since the ACP rest represents a generic rest for which the structure cannot

be represented, SMILES structures with (*) denoting the ACP rest as well as molecular formulas containing an R group for the ACP rest has been added for all intermediates. Fatty acid biosynthesis is performed until a length of 16 carbons is reached for straight-chain fatty acids and 13 carbons in the case of branched-chain fatty acids. Afterward, fatty acids are elongated in the ER. Both cytosolic and mitochondrial fatty acid synthesis has been added. Evidence exists that *C. elegans* contains a function mitochondrial fatty acid biosynthesis pathway [253]. Another problem that emerged by merging the models was mixing the different forms of ACP. Mistakenly, all reactions used the apo-ACP version, which represents the freshly synthesized protein. This protein has to be modified with a 4'-phosphopantetheine residue to form the active protein. The corresponding reactions have been added and corrected.

Table 8: Number of reactions associated with newly generated metabolic pathways related to fatty acid biosynthesis, elongation and desaturation.

Pathway	Number of reaction
Fatty acid biosynthesis (odd, iso chain, cytosolic)	19
Fatty acid biosynthesis (odd, iso chain, mitochondrial)	19
Fatty acid elongation and desaturation (odd, iso chain, ER)	41
Fatty acid biosynthesis (even, straight chain, cytosolic)	36
Fatty acid biosynthesis (even, straight chain, mitochondrial)	36
Fatty acid elongation and desaturation (even, straight chain, ER)	67

Further elongation of fatty acids follows the same principle as fatty acid biosynthesis. However, it occurs in the endoplasmatic reticulum (ER) and uses acyl-CoAs instead of acyl-ACPs. *C. elegans* harbors nine genes encoding for elongases, named *elo-1* to *-9*. They are homologous to the human ELOVL genes. These transferases catalyze the first reaction of the elongation, the condensation of the fatty acyl CoA with malonyl CoA. *C. elegans* enzymes for condensation also show substrate specificity. Palmitic acid is elongated to

stearic acid by ELO-2, while γ-linolenic acid and stearidonic are elongated to dihomo-γ-linolenic acid and eicosatetraenoic acid, respectively by ELO-1 and ELO-2 [252, 254, 255]. ELO-5 and ELO-6 are working on the elongation of monomethyl-branched chain fatty acids. For all other ELOs, the specificity remains to be elusive. The remaining step of the elongation, the 1st reduction, dehydration, and 2nd reduction, are catalyzed by the LET-767 β-keto-acyl reductase [256], and most like HPO-8 (dehydratase) and ART-1 enoyl reductase.

Fatty acid desaturation is another important reaction in fatty acid biosynthesis to produce mono- (MUFA) and polyunsaturated fatty acids (PUFA). In contrast to mammals, *C. elegans* can produce all required PUFAs on its own and does not have to take them up from food. The *fat* gene family encodes for different fatty acyl desaturases and has in total of seven members. *fat-1* encodes for an ω-3 desaturase preferentially working with length from C16 to C20. It is known to produce α-linolenic acid from linoleic acid, stearidonic acid from γ-linolenic acid, eicosatetraenoic acid from dihomo-γ-linolenic acid and eicosapentaenoic acid from arachidonic acid. Mutation of *fat-1* leads to an increase in arachidonic compared to wild-type worms [252]. FAT-5, -6- and -7 have a Δ9 desaturation activity, while FAT-3 is a Δ6 desaturase and FAT-4 is a Δ5 desaturase [257, 258]. Most important for correct metabolic modeling is the correct reaction balance. Protons and water have been added to the reaction to balance all reactions correctly.

Desaturation is an important reaction to produce unsaturated fatty acids. Interestingly, Browse et al identified alternative biosynthetic pathways in *fat-6;fat-7* double mutants [258]. An alternative pathway for fatty acid desaturation and elongation has been described by Brock et al. in *fat-6;fat-7* double mutants. In this double mutant the unusual fatty acids FA 18:1(13Z), FA 18:3(8Z,11Z14Z), FA 18:4(5Z,8Z,11Z,14Z) and FA 18:4(8Z,11Z,14Z,17Z) have been detected using GC-MS. A pathway for the synthesis of these fatty acids has been proposed. These new reactions potentially compensate for the loss of both desaturases and produce new and unknown fatty acids. These reactions have been curated, but under normal conditions, they are not active. They can be inserted if a double mutant of *fat-6* and *fat-7*

shall be studied using the GSMNs, or their flux can be set per default to 0 to not influence fatty acid production under normal conditions. These reactions should be only used if the double mutant is studied.

Besides these well-known reactions for elongation and desaturation in *C. elegans,* potential further reactions exist. Very long-chain fatty acids and 2-hydroxy derivates are found in *C. elegans* sphingolipids. Furthermore, different very long chain ascarosides with chain lengths of up to 33 carbons have been found in *C. elegans* [139]. Very long-chain fatty acids and 2-OH fatty acids are required for sphingolipid metabolism. It remains elusive at the current stage which genes are required for the biosynthesis of these long-chain fatty acids.

Several other fatty acids have been detected; for example, Gao *et al.* detected the presence of a FA(24:6) [130]. Such a fatty acid can be produced by two further elongation steps starting from eicosapentaenoic acid (FA 20:5(5Z,8Z,11Z,14Z,17Z)) and desaturation at position 6. Also, intermediates like FA(22:5) and FA(24:5) have been detected, suggesting that this pathway might be present in *C. elegans*. However, no enzymes for these reactions have been characterized in the nematode.

2.2.5.3. Fatty acid β-oxidation

Fatty acids can undergo β-oxidation upon energy demands in *C. elegans*. Reactions for both versions of β-oxidation, peroxisomal and mitochondrial, have been included in the previous *C. elegans* models and the merged WormJam model. Parts of the metabolism take place in the peroxisomes, especially the breakdown of long-chain fatty acids, while the mitochondria metabolize shorter chains, but certain overlaps in substrates exist. However, similar to the biosynthesis of fatty acids, both pathways have been curated in different detail, with several lumped reactions found. In order to improve the representation of these pathways, they have been normalized to the same level of detail as fatty acid biosynthesis. In both versions of β-oxidation, one cycle consists of four individual reactions, and all reactions have now been explicitly added with all intermediate metabolites to allow more detailed mapping.

Since *C. elegans* theoretically also contains very long fatty acids until 30 carbons. Therefore fatty acid β-oxidation also starts from C30 in the WormJam model. The degradation is split into different parts, whereas long- and very-long-chain fatty acids are degraded in the peroxisome until a length of about eight carbons. Afterward, they are shuttled to the mitochondria for further breakdown. An inevitable overlap in chain length between the peroxisome and the mitochondria exists. In order to model this correctly, an overlap of the peroxisomal and mitochondrial beta-oxidation was allowed. Fatty acids of a length from 18 to 8 carbons can be oxidized in both compartments. Iso-branched chain fatty acids are exclusively oxidized in the peroxisome producing Isovaleryl-CoA.

The individual reactions between the peroxisome and mitochondria are very similar but are catalyzed by different enzymes. The major difference is found in the first reaction. In the case of peroxisomes, this reaction is catalyzed by Acyl-CoA oxidases (ACOX). The FAD required in this reaction is recovered by converting O_2 to H_2O_2. In contrast to this, the mitochondrial version relies on Acyl-CoA dehydrogenase recycling FAD by the electron-transferring flavoprotein (ETF). All ACOX enzymes are located in the peroxisome and are also used in ascaroside biosynthesis, which also relies on peroxisomal β-oxidation. The enoyl-CoA catalyzes the second step of the β-oxidation cycle. *C. elegans* harbors different *ech* genes. In order to identify which ones are part of the peroxisomal β-oxidation the subcellular location of the protein product was inferred from the UniProt database.

Fatty acids are imported into the peroxisome as Acyl-CoA. Since the peroxisome is currently not treated as an individual cellular compartment in the WormJam, all these reactions have been added to the cytosol. In total, 92, 44 for the breakdown of saturated straight chain fatty acids from C30:0 to C10:0 and 48 for the β-oxidation of iso-branched chain fatty acids, reactions have been written to replace all old reactions related to peroxisomal β-oxidation in WormJam. The old reactions were mostly lumped reactions. They have been completely replaced with new reactions.

In the case of the mitochondria, fatty acids are imported as Acyl-Carnitines. On the cytosolic side of mitochondria, Acyl-CoAs are converted to Acyl-Carnitines by *cpt-1* (WBGene00012907), *cpt-3* (WBGene00021703), *cpt-4* (WBGene00019644), *cpt-5* (WBGene00008629) and *cpt-6* (WBGene00020911). Afterward they are imported into the mitochondria by antiport with free carnitine. A significant mistake in the merged model was the missing antiport with carnitine. From the overall elemental balance, the addition of carnitine does not change the mass balance but the cellular location of specific metabolites. The human GSMN Recon3D relies on the same reaction stoichiometry (example http://bigg.ucsd.edu/models/Recon3D/reactions/HMR_2661). These changes, therefore, make this reaction more comparable to the human version. Inside the mitochondria, acyl-carnitines are converted back to acyl-CoA by *cpt-2* (WBGene00011122).

Several fatty acids undergo this conversion in the merged model, but not all are correctly imported. Additionally, VLCFAs are also imported into the mitochondria, although they are not used there because they are oxidized in the peroxisome. Therefore, these reactions have been deleted and replaced with new reactions. Only fatty acids used by mitochondrial β-oxidation are imported by the carnitine shuttle in the updated model. To improve the import reactions for fatty acids into mitochondria reactions have been newly created. All fatty acids that are produced by the *C. elegans* biosynthesis and imported from the *E. coli* biomass have been added to undergo β-oxidation. However, based on the length, they either first undergo peroxisomal β-oxidation and are then imported to mitochondria or are directly β-oxidized in the mitochondria. In the merged model, 106 reactions are related to carnitine or the carnitine shuttle. With the corrected and cleaned version, this number was reduced to 39.

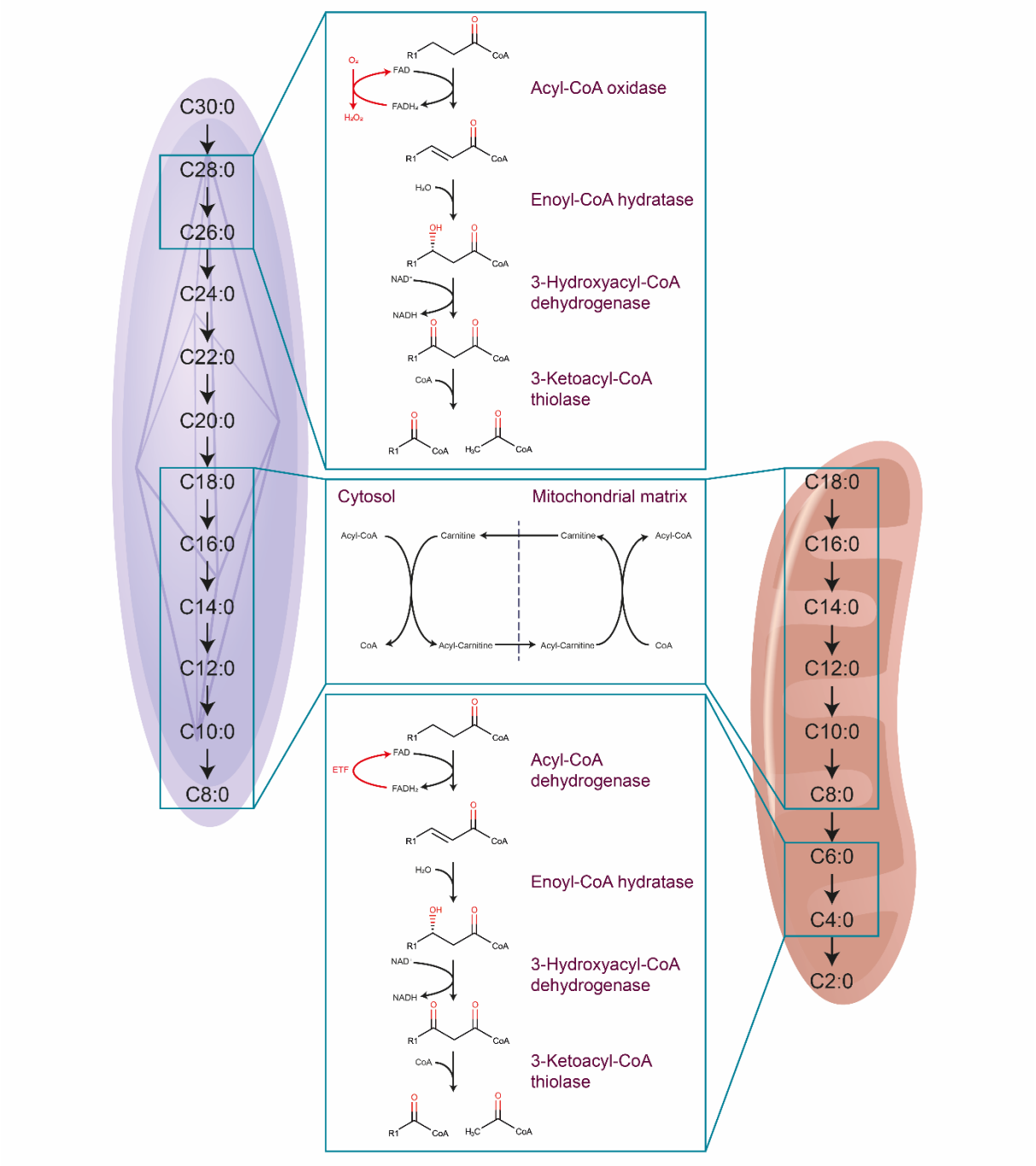


Figure 7: (A) Difference between peroxisomal and mitochondrial β-oxidation. Both reaction sequences are simliar, except for the first step. Peroxisomes use Acyl-CoA oxidases for desaturation, while mitochondria use Acyl-CoA dehydrogenases. A certain overlap in the length of fatty acids used as substrates exist. In WormJam peroxisomes break down fatty acids with a length from 8 to 30 carbons and mitochondria use 4 to 18 carbons.

The β-oxidation of unsaturated fatty acids requires additional enzymes. Double bonds in unsaturated fatty acids are typically found in the *cis* configuration and cannot be utilized by enoyl-CoA hydratases. In the case of β-oxidation of monounsaturated fatty acids, often a *cis*-3-enoyl is yielded after a few rounds of β-oxidation. This double bond is converted to *trans*-2-enoyl configuration by the action of enoyl-CoA isomerases. Reactions for the β-oxidation of

unsaturated fatty acids were found in the merged model. However, they were exclusively present in the form of lumped reactions. Therefore, individual reaction sequences were written. The β-oxidation reactions of palmitoleic (FA 16:1(9Z)), oleic (FA 18:1(9Z)), vaccenic (FA 18:1(11Z) and linoleic acid (FA 18:2(9Z,12Z)) have been added to merge model. For all these, all steps have been added explicitly, and no lumped reactions are used. At certain points, e.g. once all double bonds have been eliminated, the further oxidation of these fatty acids feeds into the normal β-oxidation or reaction sequences of β-oxidation of other unsaturated fatty acids. For example, one round of β-oxidation of vaccenic acid produces palmitoleic acid. Table 9 summarizes all reactions related to beta-oxidation that have been added. Now, these represent an accurate depiction of the metabolism of fatty acids in *C. elegans*.

Table 9: Number of reactions related to fatty acid β-oxidation. The table summarizes the final number of reactions related to the individual pathways.

Pathway	Number of Reactions
Fatty acid β-oxidation (straight chain, even, peroxisomal)	44
Fatty acid β-oxidation (iso chain, odd, peroxisomal)	48
Fatty acid β-oxidation (mitochondrial)	32
Carnitine Shuttle	39
β-oxidation of FA 18:1(9Z) (mitochondrial)	13
β-oxidation of FA 18:1(11Z) (mitochondrial)	4
β-oxidation of FA 18:2(9Z,12Z) (mitochondrial)	19
β-oxidation of FA 16:1(9Z) (mitochondrial)	13

2.2.5.4. Hydroxy fatty acids and Eicosanoids

Hydroxy fatty acids are required for different aspects of *C. elegans* biology. *C. elegans* has been shown to produce several F-series prostaglandins, which are involved in sperm guidance [259]. However, no COX or LOX homologs have been identified in *C. elegans* so

far [260, 261]. 2-hydroxylated fatty acids are part of sphingolipids in *C. elegans* [116]. The enzyme FATH-1 was shown to catalyze the reaction converting fatty acids to 2-hydroxy derivates. *fath-1* deficient animals were fed with 2-hydroxy palmitic acid in rescue experiments. Interestingly, the 2R enantiomer showed higher rescue. This is consistent with findings identifying that mammalian FA2H produces (R)-2-OH fatty acids [262]. A reaction producing (R)-2-OH fatty acids has been added to the model.

ω- and ω-1 hydroxy fatty acids are required for the biosynthesis of ascarosides. However, no CYPs responsible for the ω- or ω-1 hydroxylation have been identified so far. The *C. elegans* genome contains 97 putative CYP P450 oxidases from which one or several might be responsible for ω/ω-1 hydroxylation. To connect the biosynthesis of ascarosides (see below) with fatty acid metabolism, a mock hydroxylation reaction has been added to the model.

2.2.5.5. Ascaroside metabolism

Ascarosides are important secondary metabolites in *C. elegans* responsible for worm-to-worm communication. Chemically, ascarosides are O-glycosides of ω- or ω-1-hydroxy fatty acids and the dideoxy sugar ascarylose. Additionally, different peripheral moieties derived from different parts of the metabolism, including amino acid and nucleotide metabolism, are integrated, and modular biosynthesis of different ascarosides forms a dense network of reactions [137]. Ascarosides of different chain lengths are produced by peroxisomal ω-oxidation. The biosynthesis is dependent on different genes, e.g. *acox*, *maoc*, *daf-22,* and *dhs-28 [136]*. Different models of ascaroside biosynthesis exist and the first idea was that shortened fatty acids are attached to ascarylose. However, it turned out that the enzymes work directly on ascaroside-CoAs [263].

The ascaroside metabolism is closely related to fatty acid metabolism. Ascarosides are β-oxidized in the peroxisome, where also long-chain fatty acids are degraded. Von Reuss *et al.* have shown the presence of ascarosides with side chains of up to 33 carbons [139]. Consistent with findings by Gao *et al.*, which have detected up to C30 fatty acids in *C. elegans* metabolite extracts machinery for producing very long-chain saturated fatty acids

might exist [130, 264]. Additionally, Hänel *et al.,* a systematic study of sphingolipids has shown that they contain exclusively long-chain saturated N-acyl with chain length up to C26 [116]. *E. coli* food source is known not to produce such long-chain fatty acids. Therefore *C. elegans* has to produce these fatty acids on its own. So far, all the fatty acids found in the ascarosides were straight-chain fatty acids. So far, only very low amounts and trace levels of these fatty acids have been detected in *C. elegans* using GC-MS [105]. Therefore, all derived fatty acids might also be only present in trace amounts and are specifically synthesized in low quantities to produce the desired molecules, such as ascarosides or sphingolipids. Interestingly, ascarosides contain similar or higher levels of odd-chain fatty acids compared to even-chain fatty acids [136]. Biosynthesis of long-chain, saturated fatty acids has been added to fatty acid biosynthesis (see 2.2.5.2). However, these reactions only cover even-numbered straight-chained and odd-numbered is-branched chain fatty acids. The ascarosides contain odd-numbered straight-chain derivates, and their biosynthetic origin is not known so far.

Different blocks of reactions have been added to the WormJam model to account for the biosynthesis of ascarosides. According to the position and if the attached fatty acid has an even or odd numbered length, the reactions were split into four different pathways. Besides these basic ascarosides, complex ascarosides with additional modifications are produced by *C. elegans*. An important class of these molecules are indole-3-carboxylic acid-modified ascarosides (icas). Dependent on their length, different icas molecules have different biological activity. A short chain icas named icas#9, having five carbons in the fatty acyl induce the dauer stage and is one of the primary components of the dauer pheromone [265]. In contrast, icas#3 induces worm aggregation on food [266]. Recently the involvement of ACS-7 in the biosynthesis of indole-containing ascarosides was suggested [137]. Interestingly, the distribution of the chain length of indole-containing ascarosides does not mirror the distribution of basic ascarosides. This means that this moiety is not simply attached to existing ascarosides, but selective biosynthesis takes place. Zhou *et al.* described a biosynthetic pathway for the biosynthesis of icas#10, icas#1 and icas#9, which

has been added to the WormJam model [138]. So far, no pathways for other modified ascarosides (such as hydroxybenzoic acid modified) exist; however, isotopic labeling has shown that moieties attached to the ascarosides are produced by *C. elegans*. Since these pathways are working on CoA derivates of the different ascarosides for each step, the structures of the CoA- derivates have to be generated and added to ChEBI. Afterward, correct stochiometric versions of all reactions were generated and grouped according to the glycosidic bond and the type of fatty acid (ω or ω-1 hydroxylation). Table 10 summarizes the number of reactions added to the WormJam model.

Table 10: Number of reactions related to ascaroside biosynthesis. The table summarizes the final number of reactions related to the individual pathways.

Pathway	Number of Reactions
Ascaroside β-oxidation (even, straight chain, ascr, peroxisomal)	57
Ascaroside β-oxidation (odd, straight chain, ascr, peroxisomal)	70
Ascaroside β-oxidation (odd, straight chain, oscr, peroxisomal)	36
Ascaroside β-oxidation (even, straight chain, oscr, peroxisomal)	58
Ascaroside biosynthesis (icas biosynthesis)	13

2.2.5.6. Bile acid metabolism

C. elegans requires only low amounts of cholesterol, which so far has been suggested to have no structural function, but essential signaling molecules are derived from it, e.g. dafachronic acids [267]. These bile acid-like molecules have different roles in *C. elegans* diapause and adult longevity [131]. Although it was initially believed that dafachronic acid is a single molecule, it turned out that different structures exist, with Δ4- and Δ7-dafachronic acid representing the most studied ones. Different enzymes have been linked to the biosynthesis of the two isomers. A putative biosynthesis pathway has been described by Mahanti *et al.* [90].

Biosynthesis has two branches, one for each isomer, although interconversion of them is possible. The first step in the biosynthesis of Δ7-dafachronic acid is catalyzed by the Rieske oxygenase DAF-36, which converts cholesterol to 7-dehydrocholesterol [268]—potentially followed by an enzyme harboring Δ5 reductase activity to produce lathosterol from 7-dehydrocholesterol. Lathosterol is then converted to lathosterone by DHS-1, finally DAF-9 produces Δ7-dafachronic acid. In the second branch, cholesterol is converted to Cholest-4-en-3-one, which then gets converted to Δ4-dafachronic acid.

Reactions for the production of Δ4- and Δ7-dafachronic acids have been added. Stochiometry of all reactions was inferred from reactions annotated in SwissLipids and RheaDB. Since the exact reactions for the productions of the other dafachronic acids Δ0-DA and Δ1,7-DA are unknown, they are not curated. Likewise, the actual enzyme converting 7-dehydrocholesterol to lathosterol is not known, and therefore this reaction contains no gene association so far.

Besides the biosynthesis of dafachronic acids, the first evidence from metabolomics data that *C. elegans* produces different other bile acids exists [97, 184, 269]. Based on the homology of several enzymes to mammalian one's reactions producing these molecules have been added to the different models and also the merged model. These reactions are kept for the moment to account for the production of bile acids other than dafachronic acids.

Table 11: Number of reactions related to dafachronic biosynthesis. The table summarizes the final number of reactions related to the individual pathways.

Pathway	Number of reaction
Biosynthesis of dafachronic acids	6

2.2.5.7. Biosynthesis and Degradation of PA, DG, and TG

Lipid metabolism generally tends to be only poorly covered in GSMNs, which was also the case in the different models merged into WormJam. Biosynthesis of glycero- and glycerophospholipids starts with the transfer of a fatty acyl to the sn1 position of glycerol-3-

phosphate to form lysophosphatidic acid (LPA). In *C. elegans,* this reaction can take place in the ER or the mitochondria. The WormJam model contained only the cytosolic version (corresponding to the ER, since it is not present as an individual compartment), but not the mitochondrial. According to UniProt, products of the genes acl-4 and acl-6 are localized in the mitochondria, while the protein expressed by acl-5 localizes in the ER. The corresponding reaction for the mitochondrial versions has been added. The next step is the transfer of a second acyl group forms lysophosphatidic acid. Similar to the previous reaction, also a cytosolic and a mitochondrial version exists. The location of the mitochondrial version was wrong, and gene associations have been corrected to correctly represent the location of the different enzymes using information from UniProt. Diacylglycerol is produced from PA by LPIN-1 and can be converted back to PA by diacylglycerol kinases. Diacylglycerol lipases can produce 2-acyl-mono-glycerol, with spontaneously interconverts to 1-acyl-mono-glycerol. This molecule can be later phosphorylated to produce lysophosphatidic acid, which can re-enter biosynthesis. Furthermore, both monoacylglycerol forms can be re-acylated to form DGs. Several reactions have been deleted, which included a nuclear version of the re-acylation of MGs to DGs as well as the nuclear and mitochondrial MG-kinase reactions. For both types of reaction, no evidence exists so far in *C. elegans*. Lastly, the final acylation step to produce TG from DG has been added, and existing duplicated TG lipase reactions have been removed and replaced by a single reaction. In total, the corrected version contains 16 reactions related to the biosynthesis and degradation of PA, DG, and TG.

2.2.5.8. Biosynthesis of PC, PE and PS

The glycerophospholipids PC, PE, and PS are synthesized from DG. PC and PE are synthesized by the transfer of either phosphocholine or phosphoethanolamine headgroup from CDP-Choline or CDP-Ethanolamine to DG yielding PC and PE. Both reactions are catalyzed by the same Choline/Ethanolamine transferases CEPT-1 and CEPT-2. PS is synthesized from PC and PE by exchange of the respective headgroup moieties with L-serine. In mammals, PTDSS1, the mammalian homolog of *C. elegans* PSSY-1, accepts PC and PE as substrate, and PSSY-2 homolog PTDSS2 accepts only PE. However, if the same

substrate specificity exists for *C. elegans* PSSY-1 and PSSY-2 is unknown. The reaction related to the synthesis of PS from PE was missing and has been added. PE can also be synthesized from PS by decarboxylation. The merged model contained two versions of these reactions, one in the cytosol and one in the mitochondria, but the respective enzyme PSD-1 is known to only localize in the mitochondria. Therefore, the cytosolic version was deleted. Several wrongly annotated reactions producing phospholipids from CDP-DAG have been removed. This pathway is present in bacteria but not in animals. In total, eight reactions have been added or corrected.

2.2.5.9. Degradation and Remodeling of PE, PC, and PS

Typically, nascent glycerophospholipids are not used directly in membranes but are extensively remodeled according to the cells or organisms' needs. Fatty acids at the sn2 position are cleaved from intact lipids by the action of phospholipase A2, and the free sn2 position is subsequently re-acylated by a lysophospholipid acyltransferase. This reaction cycle is known as the Lands cycle [270]. Due to the merging process, several duplicated reactions exist also using different namespaces for the lysolipids. For example, two different identifiers have been used for putatively annotating 1-acyl-sn-glycerophosphocholine, lpchol, and ag3pc. The automatic merging process could not correctly identify this duplication and had been removed manually and several other mistakes have been corrected. According to UniProt, no phospholipase C for PCs or PEs exists in *C. elegans* so far. Therefore, all reactions related to this have been deleted. A nuclear reaction was present in the model for phospholipase D, but no evidence exists, and it has been removed from the model. Further downstream reactions hydrolyzing lysophospholipids were correctly annotated and have been renamed for consistent reaction naming. Although lipid class-specific fatty acid profiles have been generated, so far, substrate specificity for different O-acyltransferases has not been demonstrated. *C. elegans* encodes seven *mboa* genes encoding membrane-bound O-acyltransferases. Only MBOA-6 has been shown so far to incorporate PUFAs into PC, PE, and PS [271]. The total remodeling of glycerophospholipids yielded 12 reactions, which have been either corrected or newly added.

2.2.5.10. CL biosynthesis and remodeling

Reactions belonging to cardiolipin biosynthesis have been annotated as cytosolic and mitochondrial. However, the only cytosolic reaction in this pathway is the biosynthesis of CDP-DAG from PA. All other wrongly cytosolic annotated reactions have been removed. Similar to other phospholipids, cardiolipins are also extensively remodeled. No reaction that catches this aspect of cardiolipin biosynthesis has been found in the merged model. From newly synthesized CL, saturated acyl is removed by the action of a cardiolipin-specific phospholipase A2. The resulting monolysocardiolipin is then re-acylated with a PUFA to produce a mature CL. In humans, this reaction is catalyzed by tafazzin (TAZ) [272]. *acl-3* is the closest homolog that is similar to tafazzin and is located in the mitochondria and potentially catalyzes this reaction. The identity of the cardiolipin-specific phospholipase A2 remains elusive. Seven reactions specific to CL biosynthesis have been added to the model.

2.2.5.11. PI and PIP biosynthesis and degradation

The biosynthesis of PI and its phosphorylated variants forms a dense network of reactions. PI is synthesized from CDP-DAG and the sugar myo-inositol. PI shows a very specific fatty acid composition and is remodeled similarly to other phospholipids. MBOA-7 is used explicitly for the incorporation of PUFAs in PIs [151]. However, most of the reactions are associated with phosphorylation and dephosphorylation of the PI headgroup, which plays an important role in signaling, e.g. involving DAF-18/PTEN [273]. The merged model contained the mitochondrial version of reactions, which have been deleted since no evidence for them exists. Altogether, 26 reactions are related to PI and PIP biosynthesis.

2.2.5.12. Maradolipid metabolism

Maradolipids represent a novel lipid class that is found exclusively in *C. elegans* dauer larvae. Chemically maradolipids are 6,6'-diacyl trehaloses and have been identified for the first time by Penkov *et al.* [80]. Interestingly, these lipids contain a large number of BCFAs. Almost 70% of the molecular species contain at least one branched-chain fatty acid, and close to 40% two. The most abundant fatty acid was FA 15:0, most likely FA 14:0(13Me),

followed by FA 18:1 and FA 17:0. Lysomaradolipids were identified by Papan *et al.* represent potential intermediates in the biosynthesis as well as degradation products [154]. In order to incorporate the biosynthesis of maradolipids into the GSMN, putative biosynthetic reactions were added. The first reaction uses Acyl-CoA to acylate trehalose, producing lysomaradolipids, and the second reaction performs a second acylation. Similar reactions for the breakdown of maradolipids yielding a lysomaradolipid and a fatty acid have been added. Recently OAC-39 has been shown to be important for the biosynthesis of maradolipids, but it is not clear if it transfers both acyl groups [274]. Currently, four reactions are related to maradolipid biosynthesis and degradation.

2.2.5.13. Ether lipid biosynthesis and degradation

Reactions related to the biosynthesis of ether lipids in the automatically merged model contained many mistakes and were only partly annotated. First, reactions directly produced phosphatidylcholine ether lipids, and second, production of other ether lipids was completly decoupled from them. All reactions have been redesigned based on known reactions found in *C. elegans* [149, 162]. Interestingly, in comparison to mammals *C. elegans,* ether lipids consist mainly of PE species. The PC fraction contains only a minor amount of ether-linked lipids [149]. Furthermore, Drechsler *et al.* found that alkenyl bond containing PCs are virtually non-existent in the worm [162].

Several reactions were not balanced and contained major mistakes. One particular example was the fatty acyl-CoA reductase reaction producing a fatty alcohol, which was not mass and charged balanced due to the missing free CoA. Furthermore, mitochondrial version of different reactions has been found. The first step, ether lipid biosynthesis occurs in the peroxisome, and further biosynthesis is conducted in the ER. Since neither the peroxisome nor the ER is currently present in the model as distinct compartments, all reactions have been located in the cytosol, and all mitochondrial versions have been deleted. Although several enzymes of the biosynthesis of ether lipids are known, one of the key enzymes 1-

acyl-glycerone phosphate reductase converting Acyl-DHAP to LPA-O remains elusive in *C. elegans*.

The newly curated reaction sequence produces LPA-Os followed by acylation of the sn2 position to produce PA-O, followed by reactions similar to the biosynthesis of DGs and downstream phospholipids with first the dephosphorylation by LPIN-1 and the transfer of the different headgroups. First, the phosphate head group is removed to produce DG-O, and then different head groups are attached using either CDP-Ethanolamine or -Choline to produce PE-O or PC-O, respectively. Nine reactions were added. Similar to diacyl PE and PC species, lipids can be remodeled, which adds four more reactions. The exact pathways for the degradation of ether lipids are currently unknown in *C. elegans*.

2.2.5.14. NAPE and NAE biosynthesis

C. elegans uses different N-Acyl-Ethanolamides (NAE) for signaling functions, which have been shown to mediate the effect of diet on the lifespan [101]. However, so far, no complete pathway for their biosynthesis has been described. The added pathway is, therefore, hypothetical and helps to link NAE production to other metabolic pathways. Two enzymes of this pathway have been characterized. So far, NAPE-1 and NAPE-2 have been biochemically characterized as N-Acyl-Phosphoethanolamide specific phospholipase D by purification and enzymatic testing [275]. Furthermore, these enzymes are essential for axon regeneration and other processes [276]. Additionally, it has been shown that the knockdown of *faah-1*, encoding a fatty acid amide hydrolase, leads to increased levels of NAEs [101]. Still, no enzymes for the first biosynthetic steps are known in *C. elegans*. Two different biosynthetic pathways are known. While plants directly acylate the nitrogen of the PE head group using free fatty acids or acyl-CoA, animals use a transacylase which transfers an acyl from the sn2 position of a PC to the headgroup. In humans, NAEs are derived from N-Acyl-Phosphoethanolamides (NAPEs).

In the first step, an acyl group from PC is transferred to the nitrogen of PE, yielding different NAPE and LPC species. Some evidence for the existence of NAPE in *C. elegans* exists. In

their untargeted lipidomics analysis, Prasain *et al.* detected some NAPEs, although further structural characterization was not possible due to their low abundance [277]. Other parallel pathways for the synthesis of NAE exist. Besides phospholipase D, other phospholipases are able to cleave NAPEs at different positions, yielding different other molecules. Recent work from the Schroeder lab identified some glycerosphospho-N-Acyl-ethanolamides (GPNAEs), which are potentially part of a more extensive metabolic network connected to endocannabinoid synthesis [140] (Figure 8). Izrayelit *et al.* have shown that endocannabinoid and ascaroside biosynthesis are both linked to *daf-22* and *dhs-28*, enzymes from the peroxisomal β-oxidation [135].

Based on these findings, a pathway similar to the one proposed by Maccarone *et al.* for the endocannabinoid anandamide is proposed for *C. elegans* [278]. First, an acyl from the sn2 position of a PC is transferred to the PE headgroup producing NAPE. This molecule can then undergo different metabolic routes. Enzymatic activity of NAPE-1 and NAPE-2 can produce NAE directly from NAPE yielding an additional PA molecule. The activity of a putative phospholipase C yields phosphoethanolamides (PNAE) and diacylglycerols (DGs). By further enzymatic action, the phosphate group is removed, and NAE is produced. Another possible pathway removes the fatty acyl bound to the sn2 position by the activity of a PLA2 and produces lysoNAPE, on which a lysoPLD acts to produce NAE and LPA. A last alternative route removes both fatty acids from the original PE and produces a glycerophosphoethanolamide (GPNAE), from which, in successive steps, glycerol 3-phosphate is removed to produce NAE. The relation between NAE and ascarosides is currently not known. Ethanolamide containing ascarosides might be produced from ascarosides that are transferred to the PE headgroup and released upon the activity of NAPE-PLD. In total eight reactions have been added to the WormJam model. However, most of the pathways require extensive biochemical validation.

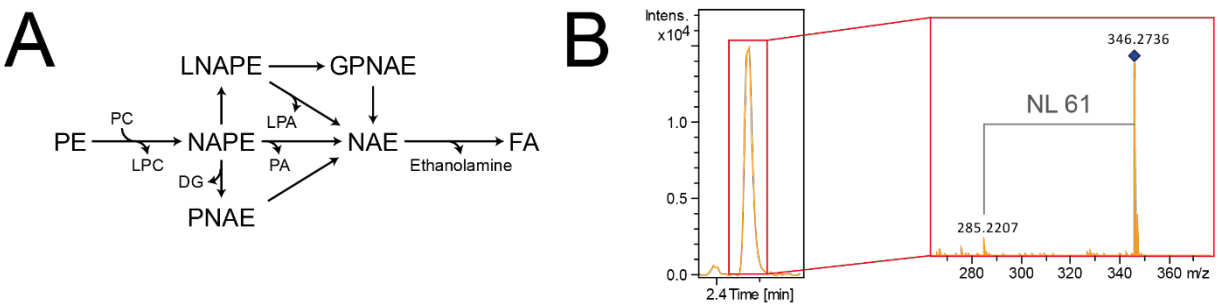


Figure 8: (A) Potential biosynthetic pathways of NAPEs and NAEs. One fatty acid side chain is transferred from phosphatidylcholine (PE) to the nitrogen in the head group of phosphatidylethanolamine (PE). Different routes to produces NAEs from NAPEs are possible. The most likely one is producing NAEs directly by means of a N-Acyl-Phosphoethanolamide specific phospholipase D. (B) Extracted ion chromatogram and tandem MS of Eicosapentaenoyl-ethanolamide (EPEA).

2.2.5.15. Sphingolipid metabolism

Sphingolipids play important roles in lipid rafts, distinct membrane areas that are required for the correct function of specific membrane-bound proteins. These rafts floating in the fluidic mosaic of the membrane and are enriched in cholesterol, sphingolipids, and saturated glycerophospholipids. However, membranes of *C. elegans* contain only low amounts of cholesterol, but sphingolipids in *C. elegans* are different from mammalian counterparts [267]. In 1995 Chitwood *et al.* showed that Glucosylceramides (GlcCer) from *C. elegans* contain a 17 carbon long and iso-branched-chain sphingoid base [279]. The same paper also has shown that N-Acyl band fatty acids contain long straight and iso-branched chained 2-hydroxy fatty acids. Interestingly, they were of odd and even numbers. The most abundant fatty acids with FA 22:0(2OH) make up 31.6 % of total GlcCer. Two years later, neutral glycosphingolipids from *C. elegans* were studied by Gerdt *et al.* [280]. In addition to iso-branched chains, they also found ante-iso-branched chain sphingoid bases. Like Chitwood *et al.*, they found that N-Acyls contain 2-hydroxy fatty acids, and again FA 22:0(2OH) was the most abundant. Glycosphingolipids of the Glc(beta)1Cer, Man(beta)4Glc(beta)1-Cer, and GlcNAc(beta)3Man(beta)4Glc(beta)1Cer series were also identified. Zhu *et al.* established a relation between the TORC-1 pathway and sphingolipids in postembryonic development of the worm [250]. Intact GlcCer are required for correct development. *elo-5* producing FA 14:0(13Me), *sptl-1* and *fath-1* were required for correct development. if one of these enzymes

was missing worms arrested at the L1 stage. Missing *elo-5* could be rescued by supply with an exogenous C17iso-branched chain sphingoid base. *fath-1* is required for the production of 2-hydroxy fatty acids and neither C15iso nor C17iso branched chain sphingoid base could rescue *fath-1* loss of function from L1 arrest. The same was true for *cgt-1* and *cgt-3* loss of function mutants. More recently, Hannich *et al.* showed that C17iso branched chain sphingoid bases are produced by *C. elegans* from Leucine using degradation products as a primer for fatty acid biosynthesis [251, 279].

In contrast to this known biology, all initial, automatically generated metabolic reconstructions contained sphingolipids with a C18 sphingoid base since gene function was inferred from homology to human genes. Based on published results, reactions have been corrected and added based on reactions supplied by the initial merged WormJam model. First, all sphingolipid structures have been corrected from C18 to C17iso sphingoid bases. Second, already in the fatty acid biosynthesis part, a reaction for the production of 2-OH fatty acids has been added (see 2.2.5.2 and 2.2.5.4). Mammals contain six different isoenzymes for the production of ceramides, named CERS1-6. All of them have a preferred substrate. The worm harbors three ceramide synthase genes, *hyl-1*, *hyl-2* and *lagr-1*. *hyl-1* and *hyl-2* are orthologs of CERS4, CERS2 and CERS5. They also show specificity towards different lengths of the acyl chains. HYL-1 produces ceramides with very long chains of >= 24 carbons, while HYL-2 prefers shorter chains with <= 22 carbons [157]. However, since reactions are only using generic metabolites for fatty acids, this behavior cannot be modeled in the model so far unless adding explicit reactions for all possible fatty acids. In total WormJam contains 28 reactions associated with sphingolipid metabolism.

2.2.5.16. tRNA metabolism

Several reactions of the tRNA metabolism were flagged to have imbalanced reactions. This was mostly due to missing formulae for the tRNA molecules. Different solutions and formulae have been proposed in BiGG. However, the solution from the RheaDB was favored since it most correctly depicted reality, using the terminal 3'-AMP residue and the respective

aminoacyl-residue as structure. tRNA-related metabolites in the model have been corrected towards this functional group, which included the addition of formulae for all tRNA and correction of all amino-acyl-tRNAs.

2.2.6. Grouping of reactions into pathways

Reactions are typically organized in pathways, which in turn can be organized in larger assemblies. In order to allow the analysis of such larger assemblies, different definitions of pathways have been added. In total WormJam contains at the moment 139 pathway definitions, e.g. “ascaroside β-oxidation (even, straight chain, oscr, peroxisomal)” or “palmitoleic acid β-oxidation”. These have been based on other pathway definitions inferred from KEGG, Reactome, and others. Reactions have been grouped as far as possible into these groups. The highest counts or reactions per pathway definition are associated with transport and exchange reactions, which do not represent true biochemical reactions but are important for metabolic fluxes across compartments. Most true reactions are associated with different parts of the ascaroside metabolism. This is due to the repetitive nature of the β-oxidation; a high number is achieved. The same is true for the biosynthesis, elongation and desaturation of fatty acids. Likewise, specific reactions are shared between different pathways. Examples are conversions of CoAs or malate dehydrogenase reactions. Additionally, to further group reactions in larger networks, super pathways have been defined. Examples of super pathway definitions are “lipid metabolism”, “fatty acid metabolism” or “ascaroside metabolism” collecting several smaller pathway definitions. These superpathway definitions are used for filtering or interpretation of data.

2.2.7. Statistics and comparison against other GSMNs

After several rounds of curation, WormJam was compared against the previous models and Worm 1.3, which represents the most current version of Worm 1 [229]. From all models, Worm 1.3.0 contained the most reactions and metabolites. This model was inferred from the previously published human reconstruction by homology. Additionally, some worm-specific reactions e.g. for some ascarosides, have been added. Despite the size of the Worm 1.3.0

model, it still contains several errors, which have been corrected in WormJam. One of the most obvious is the sphingolipid metabolism; in Worm 1.3.0 still, C18-based sphingolipid are present due to homology with human reactions. Though only a low of reactions in WormJam were inferred from the suggestion in ascaroside related papers, it represents the current best model and most complete model for ascaroside biosynthesis.

Comparing all models WormJam presents the most accurate depiction of *C. elegans* metabolism as it covers most of the *C. elegans* specific aspects. The model combines the knowledge from different models and integrates them into a consensus reconstruction. In the next steps, WormJam will be more closely integrated with RheaDB and Uniprot. Uniprot moved away from EC notation towards reaction annotation based on RheaDB reactions. First attempts for WormJam have been made to recover and compare all reactions for *C. elegans* proteins in Uniprot against current reactions. This allows assigning RheaDB reaction IDs to WormJam reactions as well as identifying gaps and potential errors in reactions. CeCon has been based on UniProt and will help in this integration. Statistics of different models are summarized in Table 12 This will be followed by a more detailed comparison against Worm 1.3.0. Since reactions are inferred from human models, a direct comparison against the human situation will be possible to identify which aspects of metabolism can be studied well. Lastly, since *C. elegans* is typically fed on *E. coli* or even a mixture of bacteria mimicking the microbiome, integration of bacterial models will be foreseeable. For the moment, WormJam serves as a reference database for metabolites present in *C. elegans*. Extensive validation of its modeling capabilities, e.g. being able to identify essential genes, etc., will be performed in the future using FBA.

2.2.8. Curation of metabolites detected in C. elegans

Currently, metabolomics can detect much more features or metabolites than metabolic models might contain, but it also often misses metabolites that are present in the metabolic modes. Different reasons for this exist. First, the choice of analytical method might not be optimal for the detection of these metabolites (e.g. RP vs. HILIC or positive vs. negative

ionization mode). Second, metabolites are not stable enough to be detected. Third, metabolites are converted with high turnover rates, making measuring them impossible. Lastly, specific metabolites simply fall below the limit of detection of the employed analytical setup.

Table 12: Statistics of different C. elegans GSMNs. WormJam shows higher numbers compared to iCel1273 or elegCyc but is still very small compared to Worm 1.3.0. Larger numbers in comparison to iCel1273 and elegCyc are explained by the number explicit reactions used in WormJam.

Category	WormJam	iCel1273	ElegCyc	Worm 1.3.0
Number of reactions	3668	1985	1914	12174
Number of genes	1301	1273	979	1732
Number of metabolites	3000	887	1640	8138

In order to get an overview of the coverage of current metabolomics approaches used in *C. elegans,* metabolites detected in the different publications have been curated. Publications were screened for detected metabolites, and a list was compiled for each one individually. In order to avoid overannotation, the closest metabolite structure has been curated. Reported identities were treated as all potentially correct without further verification. Metabolites might have been identified at different MSI levels [281]. However, these levels are often not reported. Several studies might have reported, for example, L-Tryptophan, though no chiral chromatography or analysis has been performed to prove this assumption. Nevertheless, this information was curated as, at a later stage, the comparison was only performed on the 2D structure of molecules.

In total, 65 publications have been curated. Associations between articles and metabolites have been established using the PMID or DOI as a unique identifier for the article and structural information for metabolites, which included formula, SMILES, InChI, and InChIKey, as well as database identifiers, if available. In total, 6508 associations with 3694 unique metabolites and 1399 with an explicit structure based on the presence of SMILES have been

found. The discrepancy is explained by the number of lipids curated, which often don't have an explicit structure associated with the entries due to the level of identification.

Metabolites from the curated list and WormJam were compared on different levels. First, only metabolites with an explicit structure were selected. When comparing exact matches based on the full InChIKey, 397 matches were found, while 418 matches were found based only on the first block of the InChIKey. Comparison of the first block of the InChIKey allows for ruling out ambiguity based on the stereochemistry and exact 3D structure of molecules but also collapses specific metabolites into a single entry (for example, all hexoses will have the same 2D structure).

The list of non-matching metabolites was further investigated. In the case of predicted but not detected metabolites from the WormJam model, several CoA derivates were found. The high number of CoAs is explained by different β-oxidation pathways of fatty acids and ascarosides, which contain intermediates which might be only present in very low amounts or metabolites metabolized too fast to be detected. Other metabolites that have not been detected so far include very hydrophobic substances such as retinol and derivates, ubiquinol and ubiquinones, or different steroids. These metabolites are typically detected using APCI, which is not routinely used in metabolomics. Interestingly also, a number of acyl-carnitines have not been detected so far. On the other side, metabolites detected but not present in the model so far contain several fatty acids and hydroxy fatty acids. Free 3-hydroxy fatty acids might represent breakdown products from acyl-CoAs in the biosynthesis or β-oxidation of fatty acids. Other examples are ascaroside derivatives, for which so far, no biosynthesis pathway has been added to WormJam. Likewise, several di- and tripeptides have been detected, which are non-specific breakdown products of proteins and can be, therefore, not mapped to specific entries in the model.

2.2.9. Metabolite database

In order to make all the curated metabolite lists publicly available, they have been added to a GitHub repository (https://github.com/wormjam-consortium/wormjam-db), which can be

further extended at any time. Furthermore, all scrips for data filtering and comparison against WormJam can be retrieved for (re-)analysis once new curations of metabolite lists or the WormJam model become available. So far, no *C. elegans* metabolome database exists. The curated data served as input for an initial version of such a database, which can be used for the annotation of LC-MS-based metabolomics data. All explicit and no generic metabolites are included by filtering out all compounds with an InChIKey. Since subcellular location is not required, duplicated entries have been removed. Based on the curated literature data and data from WormJam, different libraries have been generated. The first one represents the merge between both data sources and contains 2192 unique entries based on the full InChIKey. This library is used in all later chapters for the annotation of metabolites. One specialized library is the ascaroside library. It contains 462 unique metabolites in total, including CoA derivatives. SMID-DB represents an online resource of ascarosides curated by the Schroeder laboratory, but for example, does not contain the CoA derivates nor the ethanolamide ascarosides detected previously. Therefore the two libraries are complementary to each other and can be merged to yield a more comprehensive library of secondary metabolites. Likewise, not all entries in SMID-DB are related to *C. elegans*, but they have different close relatives. Lipids have been a big part of the curation of literature, but no explicit structures in the form of SMILES could be curated. However, lipid curation is discussed in the following chapter.

2.3. Detailed lipid metabolism network and lipid structure curation

The inclusion of lipids in GSMNs is a reoccurring problem. Lipids are an important class of molecules that serve as building blocks of membranes, energy storage metabolites, or signaling molecules and have to be included in the biomass equations of GSMNs for correct metabolic modeling. However, complex lipids such as PEs, PCs, etc., are typically depicted as single nodes ignoring the vast chemical complexity of this molecule class. Mostly lipid demand is constrained by the measurement of fatty acid methyl esters based on a total lipid extract. The acyl composition of all lipid classes is assumed to be the same. In order to improve the modeling of lipids, the SLIMEr approach has been developed for the yeast metabolic model [282]. This approach enhances the model by defining new pseudo reactions for the primary building blocks of lipids, the headgroups, and the acyl chains. The new model was able to predict abundances of different lipids correctly.

However, besides correct integration into biomass equations, the depiction of the exact metabolism of individual lipid species is an important aspect. Several lipid-related enzymes show specificities toward specific lipid classes and fatty acid compositions. Therefore, it is not correct to assume a homogenous distribution of acyl chains across all lipid classes. Furthermore, the possibility of mapping lipids to GSMNs is nearly impossible since most models only include generic lipid species, such as PC, and don’t consider the structural diversity. Poupin *et al.* developed a methodology to use ontologies (e.g. ChEBI ontology) to map between lipids identified in lipidomics datasets and lipid species in the GSMNs [204]. Effectively, the shortest distance between the measured lipid and the species present in the model. The closer they are in the ontology the smaller the distance.

Although this improves the matching and analysis of lipids in the context of GSMNs, it is only applicable to lipid species that can be found in the ChEBI database. So far, no exact knowledge about the number of different lipid species that can be found in *C. elegans* exists, as well is no database on *C. elegans* lipids. Different measurements of lipids in the worm have been performed, but they are currently far from being complete. An exact depiction of

lipid metabolism and all involved enzymes and lipid species is required to solve this problem. A first step has been made by improving lipid-related reactions in WormJam. However, they are still based on generic lipids. Since lipids are made of modular building blocks, lipid structures can be predicted from the combination of these different building blocks using generic reactions as a template.

In the WormJam model presented in the previous paragraphs, lipids are represented as generic nodes consuming different building blocks. These generic reactions can be used as a template to generate more specific reactions producing lipids at different levels of detail. Such a network can be used to predict the composition of individual lipidomes in different mutants or in the future, determine how fluxes in lipid pathways are changing to maintain lipid compositions. Furthermore, theoretical possible lipid structures can be predicted and used for the annotation of lipidomics data. However, detailed quantitative analyses of the *C. elegans* lipidome to compare against are still missing. Several non-targeted lipidomics analyses have been performed, and several lipids have been detected. Still, biosynthetic origin of all individual species must be connected to the rest of the metabolic pathways. Furthermore, analysis of lipidomics analysis would be significantly enhanced using lipid biosynthetic and degradation networks. The group of Michael Wakelam developed an algorithm for independent pathway analysis of lipidomics data based on the Prize-collecting Steiner tree problem on graphs [283, 284]. This method uses a given lipid metabolic network with a given lipidomics data set and extracts the optimal subnetwork [285]. Another possibility for lipid network analysis is Linex [286]. In contrast to a previously described pathway analysis, this approach takes into account the different lipid species instead of lipid classes [287].

Different tools for the creation of lipids exist, e.g. LipidCreator allows to the creation of virtual libraries of lipids and generate assays for targeted lipid analysis or spectral libraries for identification [288]. SwissLipids is based on the combinatorial generation of lipids. Six hundred twenty publications have been curated to generate a comprehensive knowledge on

lipid metabolism. Data was primarily based on data from human, mouse, *C. elegans,* and yeast. Based on the obtained lipid-related reactions and feasible fatty acids lipid structures were created. Furthermore, lipids are arranged in a hierarchical structure representing the different levels of identification [289]. However, it does not allow the flexible adaptation to create new lipidomes on the fly, e.g. removing enzymes to simulate mutations or changing fatty acid input.

The goal was to develop a flexible framework derived from the available knowledge on the biosynthesis of lipids from different sources. WormJam was leveraged to predict lipid species potentially present in *C. elegans*. The developed codebase can be used not only for the prediction of *C. elegans* lipids but also for other species. Generic templates have been generated based on reactions found in WormJam and RheaDB. Since UniProt uses Rhea reactions for functional annotation of enzymes, potential reactions can be extracted by comparing entries for a specific organism and retrieving the respective RheaDB reaction identifiers. In the case of *C. elegans,* this information is retrieved from the gene identifiers in the WormJam model and UniProt.

2.3.1. Template-based lipid reaction and structure generation

Following the example of SwissLipids, lipids that might be present in *C. elegans* are going to be generated *in silico* based on reactions present in WormJam. Instead of using plain chemical combinatorics, the creation is guided by the lipid-related metabolic pathways present in RheaDB. Reactions in metabolic models contain generic metabolites, e.g. M_fataccoa_c can be any Acyl-CoA. This generic metabolite can be exchanged for a more specific metabolite, e.g. oleoyl-coenzyme A (CoA(18:1(9Z))). Replacing all generic metabolites in all lipid biosynthetic reactions creates a multiple of new reactions. For lipids, the accepted shorthand nomenclature by Liebisch *et al.* is used [207]. This shorthand nomenclature is well suited for this purpose because it allows the identification of lipid building blocks within the notation. The individual building blocks can be isolated and used to create new lipids.

The reaction below is the SBTab representation of the glycerol-3 phosphate acyl transferase reaction.

```
M_glyc3p_c + M_fataccoa_c <=> M_coa_c + M_1acglyc3p_c
```

`M_glyc3p_c` represents glycerol 3-phosphate and `M_coa_c` free coenzyme A, both with fixed structures. `M_fataccoa_c` and `M_alpa_pl_c` represent the variable substrate and product. If this reaction were expressed as a distinct reaction using an oleoyl-residue as an example, it would look as follows:

```
Glycerol 3-P + CoA(18:1(9Z)) <=> CoA + PA(18:1(9Z)/0:0)
```

Reaction templates have been made independent of the BiGG metabolite ID namespace for easy curation. Furthermore, RheaDB differentiates between different possible directionalities for reactions: undefined, left to right, right to left, or bidirectional. This is also represented in different template reactions, each one associated with the respective RheaDB ID. To differentiate the reaction formula, =, =>, <=, or <=> are used. The reaction templates for the example above look like following:

```
Glycerol-3-P + AcylCoA = CoA + LPA
Glycerol-3-P + AcylCoA => CoA + LPA
Glycerol-3-P + AcylCoA <= CoA + LPA
Glycerol-3-P + AcylCoA <=> CoA + LPA
```

Based on the “template reactions” with generic metabolites, more detailed versions of the complete lipid metabolism can be generated. In order to correctly identify lipid species, shorthand notations, according to Liebisch *et al.* were used [207]. To not have to generate all possible reactions manually, an R-based workflow was created that replaces generic metabolites with specific ones. As a starting point, lipid reactions from WormJam were used, and the respective RheaDB reactions have been curated and templates generated.

In several instances, different building blocks need to be isolated from lipid abbreviations. Liebisch annotation made it possible to generate different regular expressions for this task.

The following nomenclature rules were used for acyl groups and sphingoid bases. The longest carbon chain and double bonds are used as the basis, e.g. 18:0. The number of carbons is separated by ":" from the number of double bonds. If double bonds are present, they are encoded with the position and their stereo chemistry, e.g. 9Z. For example, an oleoyl moiety would be encoded as 18:1(9Z). Functional groups are used in the following order: double bond (DB) > OH > OOH > O > NH2 > Me. They have to be encoded with the position, and if the atom is a stereocenter, the configuration has to be added with squared brackets. For example, 2-hydroxy palmitoyl moiety would be 16:0(2OH[R]). Individual functional groups are separated from each by a comma ",". Since the prediction works on exact structures, full details for all lipids are given. The coding of sphingoid bases follows the same rules. For example, 16:1(4E,1OH,3OH[R],2NH2[S],15Me) is the predominant sphingoid base in *C. elegans*. Lipid classes are based on the accepted code used by different databases, e.g. LipidMaps or SwissLipids, e.g. phosphatidyl cholines are abbreviated as PC. In the simplest case of reaction creation, the backbone remains the same; only the lipid classes need to be changed, for example, when creating Acyl-CoAs from fatty acids. However, if specific groups have to be extracted, regular expressions are used. Acyl groups and sphenoid bases are isolated using the same regular expression

```
"(O-|P-
)*\\d+:\\d+(\\((\\d*(E|Z|Me|OH|OOH|O|NH2|delta)(\\[(S|R)\\])*,*)*\\)
)*"
```

Since functions for the isolation of functional groups are of general interest beyond the template-based reaction generation, they have been externalized to a package called `lipidomicsUitls`, available from GitHub (https://github.com/michaelwitting/lipidomicsUtils). In addition, all generated functions have been bundled into an R package called `LipidNetworkPredictR`. This package is available for GitHub (https://github.com/michaelwitting/LipidNetworkPredictR).

2.3.2. Substrates for generation of networks

Major substrates for the generation of lipids are fatty acids in the form of acyl-CoAs. The combinatorial space of lipids is defined by the fatty acyls used by the organism at a given time point for *de novo* synthesis or remodeling of lipids. In order to get an impression of fatty acids found in *C. elegans,* and (potentially) bound in different lipid classes, data from several publications analyzing the fatty acid content of *C. elegans* has been collected. Fatty acids have been categorized into saturated, monounsaturated, polyunsaturated, monomethyl branched chain and cyclopropane fatty acids. Several publications describe the analysis of either total fatty acids or lipid class-specific profiles typically determined by GC-FID or GC-MS. Figure 9 shows profiles selected from different publications [105, 147-149, 290-292].

Several fatty acids are directly taken up from the food bacteria, while others are synthesized entirely by *C. elegans*. In order to estimate the significance of the amount of fatty acid, an arbitrary cut-off of 5% of total fatty acids was used. Fatty acids below 5% were considered minor species, while above 5% they were considered major species. While the *E. coli* diet of *C. elegans* is rich in the saturated fatty acids lauric acid (FA 12:0), myristic acid (FA 14:0), and palmitic acid (FA 16:0), they are only minor species in *C. elegans*. Interestingly, palmitic acid and stearic acid have high percentages in PIs [151]. The nutritional fatty acids cis-9,10-methylene hexadecanoic acid (FA 17:0[9-10cy3:0]) and cis-11,12-methylene octadecanoic acid (FA 19:0[11-12cy3:0]) are enriched in triacylglycerols compared to other lipid classes but are also found to a certain extent in other lipid classes. Some differences between lipid classes in their fatty acid composition exist. PC and PE are the primary building blocks of membranes, but they differ in their fatty acid composition. Significant differences are found in several saturated and polyunsaturated fatty acids. While PCs only contain 2.17% palmitic and 2.61% stearic acid, PEs contain 6.43% and 11.43%, respectively. A further striking difference is the amount of eicosapentaenoic acid. While PC contains 33.58%, PEs only contain 12.60% [149].

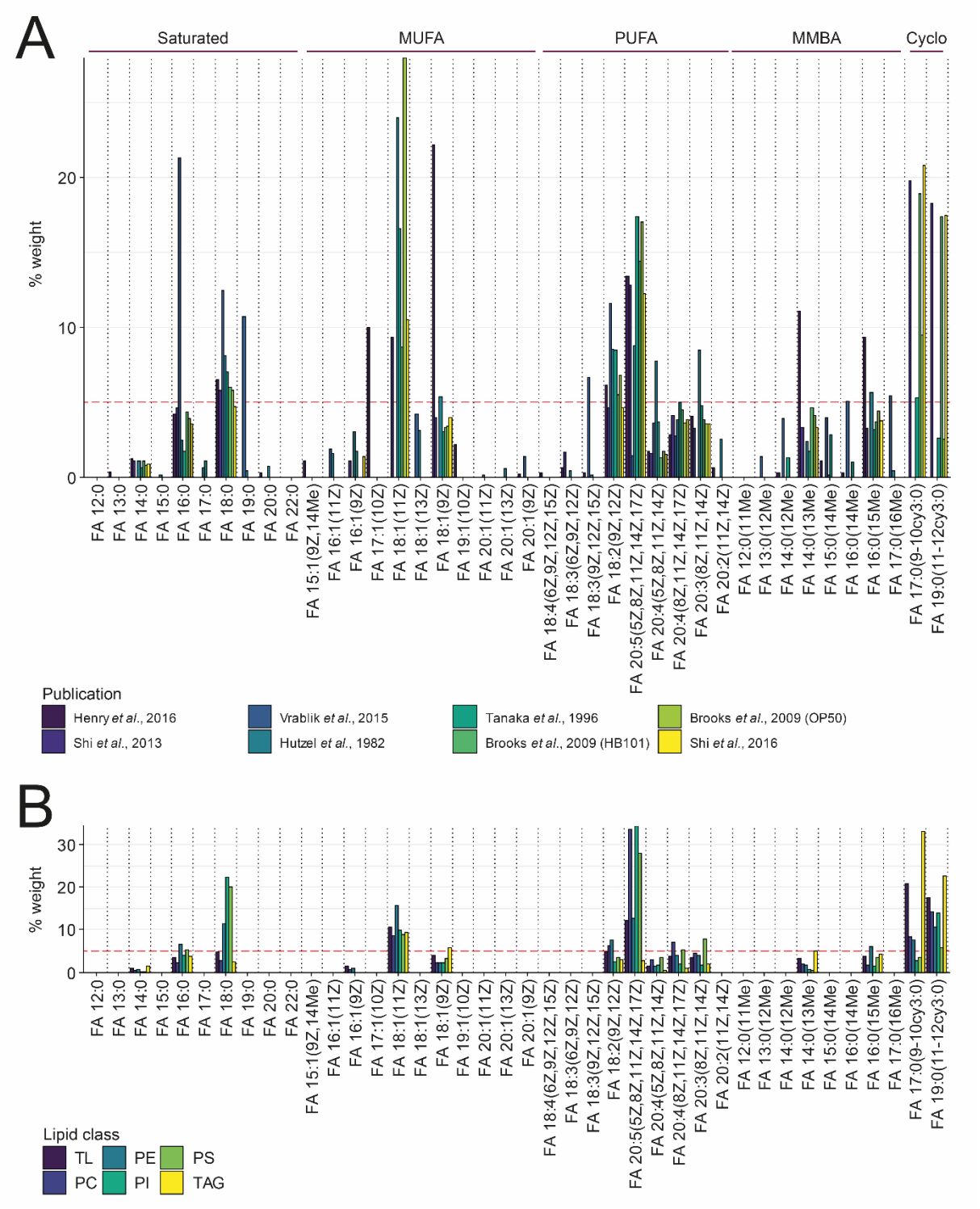


Figure 9: (A) Fatty acid profiles from different selected publications as established by GC-FID or GC-MS using saponification. Fatty acids are grouped according to their classification into saturated, monounsaturated, polyunsaturated, monomethyl branched or cyclopropane fatty acids. (B) Fatty acid profiles of different lipid classes as measured by Shi et al. [149].

Branched-chain fatty acids 13-Methylmyristic acid (FA 14:0(13Me)), and 15-Methylpalmitic acid (FA 16:0(15Me)) are exclusively produced by *C. elegans* [123]. However, in most fatty acid profiles of most lipids classes, they were only found as minor species. One exception is

maradolipids exclusively found in dauer larvae. FA 14:0(13Me) plays an important role as a precursor for sphingoid bases in *C. elegans* sphingolipids.

The fatty acid profile of sphingolipids shows different patterns. Typically, mild alkaline saponification is used to generate free fatty acids, which are in turn analyzed as their methyl esters by GC-MS or direct transmethylation followed by GC-MS. Amide-bound fatty acyls in sphingolipids are inert under these conditions and require a different strategy. Therefore, total fatty acid profiles normally represent ester-bound fatty acids ignoring fatty acyls from sphingolipids. Chitwood *et al.* and Gerdt *et al.* performed an analysis of glucosylceramides in *C. elegans* and identified several 2-OH fatty acids, which are not found in glycero- and glycerophospholipids [279, 280]. These fatty acids are long-chain saturated fatty acids ranging from 16 to 26 carbons and also include odd-numbered chains. Interestingly, also even numbered iso branched-chain fatty acids were reported.

Based on the list of detected fatty acids from the different discussed publications, a list that is used for *in silico* lipid network and structure generation was built. Shorthand notation by Liebisch *et al.* has been used for this list and all further downstream lipids. This list contained fatty acids, which are part of the glycero- and glycerophospholipids. For the generation of sphingolipids, a second list has been compiled. This list is based on the description of fatty acyl distributions by Gerdt *et al.* and Chitwood *et al.* and includes several different possible candidates for acyl groups based on fatty acid biosynthetic pathways, including 2-OH modified fatty acids.

2.3.2.1. Draft lipid network

Based on the generated list of fatty acids, a draft lipid network is created using the two different fatty acid pools for glycero- and glycerophospholipids and sphingolipids. In order to store as much chemical information as possible `rgoslin` has been used for parsing of lipid shorthand notations [293, 294]. The output delivers a data frame containing the shorthand notation on different levels as well as the sum formula and buildings block, which can be used for the prediction of MS^2 spectra, for example.

In the first draft of the network, 38 fatty acids were used as input for the biosynthesis of glycero- and glycerophospholipids and 56 fatty acids for the biosynthesis sphingolipids. No constraints were applied to the generation of reactions and lipids. 2161280 lipid structures were predicted using 38 and 56 fatty acids, respectively, and 61 reaction templates as input. The model yielded 2299902 reactions in total. This initial model ignored the subcellular localization of enzymes and lipids, including their site of production. The intention was to generate a list of possible theoretical lipids that can be used for the annotation of lipidomics experiments in *C. elegans*. No specificity in side chain position or preferential side chains in different lipid classes was included. However, the model was able to generate a comprehensive list of lipid structures and their biosynthetic routes.

Several reactions include the transfer of certain groups from one lipid class to another. In the simplest case, this group is transferred from a single metabolite, e.g CDP-Choline transfers a phosphocholine headgroup. In some other cases, a group might be transferred from different lipids, e.g. migration of acyl chains or the transfer of the phosphocholine headgroup from PCs to Cers to produce SMs. Here, the exact lipid species that transfers the headgroup is not known, and it is possible that multiple lipid species are accepted as a donor.

The produced network is much larger than most GSMNs and, therefore will be hard to model. Some possibilities to reduce the number of reactions exist, such as reducing the number of substrates for reactions by constraining the fatty acyls used as input.

2.3.3. Constraining the lipid reactions

Using the developed workflow, a large number of possible lipid species from different classes were generated. Since reactions were unconstrained, potentially several lipids that are not produced by *C. elegans* have been generated. Currently, little is known about the substrate specificity of *C. elegans* lipid enzymes. One exception is *mboa-7* which has been shown to incorporate PUFAs into PIs [151].

At the current stage, the method considers no specificity of enzymes, and fatty acyl residues are equally distributed across all classes. Hence all lipids have the same probability of being

synthesized. However, measurements show that each lipid class has a distinct fatty acid profile.

A very extensive combinatorial explosion is yielded for CL species. Using the unconstrained model, 2,085,136 different CL species are predicted. This is way over what has been currently detected in *C. elegans*. CL species have been seen in several lipidomics applications in *C. elegans* [114, 163, 164, 295]. The composition of *C. elegans* CLs is shown to be different from other species. Recently, the structural diversity of cardiolipins across different organisms has been investigated [296]. At the same time, the most abundant fatty acyl species in the different organisms were FA 16:0, FA 16:1, FA 18:1, FA 18:2, and FA 18:3, similar to humans with FA 18:2, one of the major species. *C. elegans* shows high levels of FA 20:3, FA 20:4 and FA 20:5. In all *C. elegans* studies, CL(80:15) was the highest abundant species. Results suggest that the production of CL species can be constrained towards species containing several poly-unsaturated fatty acids. Investigating the fatty acyl groups that were identified, it would be determined that most species contain C18 and C20 unsaturated fatty acids. Therefore, reactions were constrained towards PG and CDP-DG species containing only these fatty acyls, as they represent the most representative ones. However, CLs are typically extensively remodeled, and to what extent remodeling takes place in C. elegans and which species of PGs and CDP-DGs are preferred for the synthesis of nascent CLs remains elusive. Using the constraints led to a roughly 40-fold reduction of CLs species.

Another possibility for constraining reactions was identified for ether lipids. Drechsler *et al.* and Shi *et al*. analyzed ether lipid metabolism and showed that they mostly contain alkyl and alkenyl groups of 18 and 20 carbons [149, 162]. Therefore, reactions of ether lipids have been constrained to only produce lipids with 16, 18, and 20 carbons at the sn1 position. This reduced ether lipids for all classes from 1444 species to 114.

Even with constraining, in total, 161700 lipids were predicted. This number is still highly dominated by complex species such as TGs and CLs. In a later chapter, predictions will be

compared against measurements from the different measured projects as well as lipids curated from the literature. Run times for the prediction are remarkably long for lipid classes with multiple acyl chains, such as TGs or CLs. However, predictions can be performed within less than half a day. The best possibility for an increase of speed is the restriction of possible substrates and constraining of reactions.

2.4. Conclusion

GSMNs are important tools for better understanding of metabolism. They integrate all the current knowledge about the metabolism of a specific organism. For *C. elegans,* different GSMNs are available and the WormJam model represents one of the best-developed models in terms of the accuracy of the nematode's biochemistry.

Metabolomics data can help to complement findings based on GSMNs and allow to expand models by adding previously missing metabolites. However, the overlap between metabolites predicted by WormJam and measured in different publications is relatively low. Especially a lot of metabolites have been detected which couldn't be mapped to the metabolic networks. Various possibilities exist, e.g. wrong annotation or missing metabolic pathways in the model. Refinement of models such as WormJam needs to follow current metabolomics approaches in the future. Likewise, resources such as GMSNs can be used as a first-line database for the annotation of metabolomics data.

A major blind spot of GMSNs, not only WormJam, are lipids. In most models, complex lipids are treated as single entities, which does not reflect real life. An approach to creating lipid-specific reaction networks was developed to predict lipid species potentially present in an organism and their associated lipid metabolic pathways. However, the number of predicted reactions outnumbers even the largest GSMNs, which is not realistic to be used in modeling approaches. These lipid models need further refinement to be of greater use. Lipidomics in *C. elegans* will also help to develop these models further.

The usability of WormJam and the metabolites and lipids predicted will be demonstrated in the following chapters. All generated databases serve as input for annotation purposes and allow to link of metabolites directly to their respective pathways. Upon constant improvement of the WormJam GSMN as well as the developed lipid prediction workflow significant knowledge gaps, especially for the biosynthesis of secondary metabolites, will be closed. Lastly, it is aimed to make all the collected information more accessible to other *C. elegans* scientists, e.g. designing a specific webpage similar to the HMDB.

- This page is left intentionally blank -

3. Metabolite and Lipid Identification

Chapter-related publications:

N-Alkylpyridinium sulfonates for retention time indexing in reversed-phase-liquid chromatography-mass spectrometry-based metabolomics

Stoffel R., M. Quilliam, N. Hardt, A. Fridstrom, M. Witting

Analytical and Bioanalytical Chemistry, 2021 Dec 15. doi: 10.1007/s00216-021-03828-0

This article presents the initial description of an LC-MS retention time indexing system for reversed-phase separation. This system is based on the use of N-Alkylpyridinium sulfonates developed by Dr. Michael Quilliam. Together with Dr. Quilliam I discussed the use of this system in metabolomics. In this work, I created the study design and supervised Rainer Stoffel during his bachelor thesis performing all measurements. In addition, I analyzed the obtained data.

UHPLC-IMS-Q-ToF-MS analysis of Maradolipids, found exclusively in Caenorhabditis elegans dauer larvae

Witting M., U. Schmidt, H.-J. Knölker

Analytical and Bioanalytical Chemistry, 2021 Mar;413(8):2091-2102. doi: 10.1007/s00216-021-03172-3

Ion mobility is a powerful tool that can be used to identify lipids. In this article, I performed the analysis of *C. elegans* dauer larvae and identified different maradolipids. I completed the growth, extraction, and analysis of the *C. elegans* samples and the measurement of the reference standards synthesized by Dr. Ulrike Schmidt and Prof. Dr. Hans-Joachim Knölker.

LipidFrag: Improving reliability of in silico fragmentation of lipids and application to the Caenorhabditis elegans lipidome

Witting M., C. Ruttkies, S. Neumann, P. Schmitt-Kopplin

PLoS One. 2017 Mar 9;12(3):e0172311. doi: 10.1371/journal.pone.0172311

This article describes the adaption of MetFrag for the annotation of lipids using *C. elegans* as an example. In this work, I have performed the measurement of reference lipid standards and materials as well as the extraction and analysis of *C. elegans* samples, and the initial annotation of lipids based on manual interpretation and data analysis.

The metaRbolomics Toolbox in Bioconductor and beyond

Stanstrup J., CD. Broeckling, R. Helmus, N. Hoffmann, E. Mathé, T. Naake, L. Nicolotti, K. Peters, J. Rainer, RM. Salek, T. Schulze, E. Schymanski, MA. Stravs, EA. Thévenot, H. Treutler, RJM. Weber, E. Willighagen, M. Witting, S. Neumann

Metabolites. 2019 Sep 23;9(10). pii: E200. doi: 10.3390/metabo9100200

This review describes metabolomics-related packages in the R programing language. The idea to collect all available packages started in 2016 during a workshop I organized with Dr. Steffen Neumann and Dr. Jan Stanstrup at the metabolomics conference in Dublin. In this review, I helped to collect packages related to the annotation and identification of metabolites from LC-MS data.

A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R

Rainer J., A. Vicini, L. Salzer, J. Stanstrup, J. M Badia, S. Neumann, M. A Stravs, V. V. Hernandes, L. Gatto, S. Gibb, M. Witting

Metabolites. 2022 Feb 11;12(2):173. doi: 10.3390/metabo12020173.

This article describes the R packages `MetaboCoreUtils`, `MetaboAnnotation`, and `CompoundDb`. I have been involved in developing the functions in `MetaboCoreUtils` and `MetaboAnnotation` for annotating MS^1 and MS^2 data and have written all related functions and unit tests.

3.1. Introduction

Metabolite or lipid identification represents one of the major bottlenecks in non-targeted metabolomics and lipidomics. Different analytical technologies deliver different degrees of information for identification. The LC-MS-based non-targeted analysis creates a matrix containing RT – *m/z* pairs as well as the respective intensity or peak area for each sample. Without identification, this matrix contains only a collection of numbers without biological meaning. To derive biological knowledge, these numbers need to be translated into biological entities such as metabolites. Ultimately identified metabolites can be mapped to biological pathways, and knowledge about pathway activity and regulation can be derived.

Metabolite identification uses different types of information, which includes MS^1 (isotope pattern and adducts), MS^2 to MS^n (fragmentation pattern), retention time (RT) and ion mobility data as well as NMR data of purified metabolites, if possible. The different level of information allows deducing different levels of identification ranging from sum formulae to complete structures, dependent on the methods employed. The most accurate identification is achieved by comparison of measured signals with a chemical reference standard under identical analytical conditions. For correction identification, two independent physicochemical properties must match (e.g. RT, MS^1 and MS^2). This is referred to as level 1 identification according to the Metabolomics Standard Initiative (MSI) identification scheme [281]. Lower levels rely on matching fragmentation or NMR spectra against spectra from a different database, not measured under the same conditions (level 2). All other levels are referred to as putative annotations. It must be mentioned that annotation is not identical to the identification. Although this scheme partially enabled the judgment of the “goodness” of metabolite identifications, it still has several shortcomings. For example, D- and L-Tryptophan can have the same RT, m/z, and fragmentation pattern on achiral chromatographic separation. If L-Tryptophan was measured as standard and compared against the unknown substances and parameters matches, it might be reported as L-Tryptophan. However, as long as no chiral analysis was conducted, this result is overreporting. Therefore, the scheme might be partially misleading. Several other schemes

have been invented, e.g. by Schymanski *et al.* [297]. This scheme was initially described for the identification of substances in the non-targeted screening of environmental samples but is also more and more applied in metabolomics. Typically, only metabolites on level 1 are called identified, while the lower levels refer to annotated metabolites. In this work, the levels according to Schymanski *et al.* as summarized in Table 13, are used.

Table 13: Confidence in identification of metabolites can be reported by using the level system developed by Schymanski et al. [297]. While levels 2 to 4 are only called annotation, metabolites on level 1 are referred as identified. The level for metabolite annotation/identification depends on the experimental effort and the information used in the process. Level 1 identifications require a reference standard and are normally only achieved for a low number of metabolites.

Level	Identification confidence	Minimum data requirements
1	Confirmed structure	MS^1, MS^2, RT, Reference Std.
2a	Probable structure by library spectrum match	MS^1, MS^2, Library MS2
2b	Probable structure by diagnostic evidence	MS^1, MS^2, Exp. data
3	Tentative candidate(s) Structure, substituent, class	MS^1, MS^2, Exp. data
4	Unequivocal molecular formula	MS^1 isotope/adduct
5	Exact mass	MS^1

An essential question in metabolite identification is the relevance of the identification for the organism studied. Different databases contain a different number of metabolites, but not all metabolites might be present in all organisms, which is especially true for metabolites from secondary pathways. Metabolite identification in other organisms, therefore highly relies on the availability of curated metabolite databases for this specific organism. A major example is the Human Metabolome Database, which serves as a reference for human metabolomics investigations [86, 87, 298]. This database not only stores metabolites but also has a rich metadata section associated with each metabolite. This includes information on

concentration in different specimens, e.g. plasma or urine. Such databases are missing for many organisms. In the case of *C. elegans* so far, no specific database has been described. Therefore, the WormJam model described in chapter 2 serves as the first instance of a *C. elegans* Metabolome Database. While HMDB contains MS2 and NMR spectra of several metabolites, such data is not available for *C. elegans*-specific metabolites. Only recently, SMID-DB.org added MS^2 spectra for *C. elegans* secondary metabolites such as ascarosides [140]. However, these spectra are not available in an open format for metabolite annotation at the moment.

Besides organism-specific metabolite and metabolic pathway databases, reference databases for tandem MS and/or NMR data are required. Different databases are available, e.g. MassBank, MassBank of North America, GNPS, Metlin, and others [299-301]. These databases can help to identify metabolites by comparison of measured spectra against spectra stored in the different databases.

3.1.1. Metabolite and lipid identification workflows

Identification of metabolites and lipids needs to integrate as much available data as possible. Below, the information content that can be derived from each type of data is discussed. Typical metabolite identification workflows start from annotating MS^1 data with putative metabolites and sum formula calculation from isotope patterns. Based on these, further steps are then undertaken. In most cases, dereplication is performed, which means that metabolites and lipids that have been previously identified are annotated and identified. This is typically accomplished with an in-house database that contains *m/z* values, retention times, and tandem MS spectra of standards measured under identical conditions to the sample. Different properties of molecules measured by LC-MS can be used for the identification of parts of the identification.

3.1.1.1. MS^1

The information from the MS^1 level includes information of the type of adduct, the isotopes and also on potential in-source fragments. The isotopic pattern can be used for formula

calculation. Based on the resolution and mass accuracy of the employed MS analyzer, the resulting list might differ in length. Likewise, from a molecular sum formulae the isotopic pattern can be predicted. It is understood that the accurate *m/z* of a small molecule is insufficient for its structural elucidation following the example of Kind and Fiehn: Searching ChemSpider with *m/z* 378.1678 Da and ten ppm mass accuracy returns more than 9,500 structures, but even searching the exact molecular formula $C_{20}H_{26}O_7$ results in 300 possible structures [47]. Hence, the accurate mass of a small molecule cannot provide information beyond its potential molecular formula. Trying to identify a small molecule based on its mass will result in a long list containing the putatively correct identity and numerous false annotations. Also, especially in the case of lipids, potential isobaric overlap, e.g. between $[M+H]^+$ and $[M+Na]^+$ adducts, exist, which can be only resolved by ultrahigh resolution MS. Fiehn and Kind have shown that at higher *m/z* range even one ppm mass accuracy is not sufficient for correct calculation of sum formulae, however, if 5% error in the isotopic pattern is allowed results are sufficiently unique [47]. The most important factor in avoiding a combinatorial explosion of possible sum formulae is the restriction of possible elements.

With the advent of ultrahigh-resolution MS instrumentation isotopic fine structures are now within reach. While older or lower resolution instruments do not differentiate between the different possible isotopic species, the latest generation of Orbitrap and ICR-FT-MS can be distinguished if a peak is related to ^{13}C or ^{15}N or $^{13}C_2$, or ^{34}S, for example. This accuracy allows for determining elemental ratios directly from detected isotope patterns [302]. Together with the seven golden rules stated by Kind and Fiehn, an easy calculation is possible [303]. Although the current generation of time of flight (ToF) instruments does not allow this degree of detail, the obtained data is sufficiently good to narrow down the list to a reasonable number of candidates. The list of potential sum formulae can be further reduced by incorporating MS^2 data. This is done e.g. by Bruker SmartFormula 3D or Sirius [304].

Information content of the MS^1 level is limited since several structures can have the same molecular formula and are dependent on the resolution of isobaric overlap. However,

network approaches such as NetCalc or MetNet can help group *m/z* values or chemical formulae into a biological context [305, 306]. Both tools follow the idea that a (bio)chemical transformation is associated with a difference in mass between two related metabolites, e.g. elongation of fatty acids would lead to a mass difference of 28 Da or C_2H_4 as a formula. If exact *m/z* are measured on a high-resolution MS (HR-MS), then exact mass differences can be calculated and compared against a list of known mass differences from (bio)chemical transformations. NetCalc uses this information together with seed formulas of known metabolites for the calculation of sum formulae. MetNet combines mass differences with other metrics, such as correlation analysis, to derive meaningful biochemical networks from MS data. For a more comprehensive overview, see Amara *et al.* [307].

3.1.1.2. MS^2

MS^2 information can also be used to inform sum formula calculation. In the case of exact mass, MS^1 and MS^2 sum formulae can be calculated for each fragment, which must represent a valid sub-formula of the MS^1 formula and potentially can be explained by a specific neutral loss formula. The Sirius software uses the MS^1 and MS^2 information and generates fragmentation trees with sub-formulae for each possible fragment [304, 308]. By this integration, typically, a higher number of correct sum formulae are achieved. However, formula calculation is becoming more problematic for molecules > 500 Da due to the large combinatorial explosion of elements. ZODIAC has been recently integrated into the Sirius software and uses Gibbs sampling for molecular formula annotation based on the idea that metabolites do not represent isolated identities but are part of metabolic networks [309]. Molecular formulas are reranked by consideration of joint fragments and losses between fragmentation trees of metabolites detected within the same dataset. This improves correct molecular formula identification because similar structures should produce similar fragments and, therefore, similar fragmentation trees.

MS^2 can provide information on substructures present in a molecule. Specific neutral losses and fragmentation rules can be used to identify structural motifs. Different databases contain

a myriad of MS^2 spectra for other metabolites. The most used MS^2 databases are Metlin, MassBank, MassBank of North America, and GNPS libraries [299, 300, 310, 311]. Compared to structural databases such as PubChem, these databases are several orders of magnitude smaller. Furthermore, the chemistry of molecules covered is somewhat limited. Most laboratories that build and share mass spectral libraries focus on metabolites of interest for them and compounds commercially available. Several metabolite substance libraries have been commercialized in recent years, but often contain only metabolites that are cheap and stable. Although shared spectral libraries increase in the number of spectra, the number of new compounds only grows slowly. Though spectral libraries are considered as “gold standard” in identification, there are several problems associated with them. First, often the exact instrument used is not covered by public libraries. Differences in instrumentation can cause slight or large differences in spectra. In most cases, the collision energies are calculated on a different basis between different vendors. Ramped or stepped spectra, which collate multiple collision energies, allow much better comparison and often also contain more information compared to single collision energies.

Besides library matching, different *in silico* tools have been generated to speed up initial annotation and to narrow down potential candidate lists. Among the most used tools are MetFrag, CSI:FingerID, and CFM:ID [312-314]. For example, their performance has been evaluated in the CASMI 2016 contest [315]. The advantage of these tools is that they only rely on spectral data for training but allow the search in structural databases, which are much larger afterward. Recent developments allow tools to rival library matching, e.g. the COSMIC confidence score allows the separation of true from bogus annotations [316].

Another possibility for the analysis of MS^2 is the construction of molecular networks, e.g. using Global Natural Products Social Molecular Networking (GNPS) [317]. Here spectra similarities are calculated based on different scores to infer chemical similarity. Similar structures yield similar fragments or contain the same neutral losses. Following this idea, spectral similarity should be a proxy for structural similarity. However, while similar structural

parts would yield the same fragments, another variable part could lead to different fragments, e.g. different lengths of side chains. To account for this fact, the similarity measure in GNPS uses the mass difference between the precursor *m/z* of the two spectra to be compared. Fragments are considered matching between the spectra if they either match directly within a specific mass error or the difference between the precursor *m/z* values and a specific mass error. Different other tools have been built around this approach and enable in-depth elucidation of metabolomics data, e.g. MolNetEnhancer [318].

In the case of *C. elegans* no comprehensive MS^2 spectral database exists. Most central metabolites from the WormJam model that at least one MS^2 spectra in any of the above-mentioned databases,s but secondary metabolites like the ascarosides are entirely missing from these databases. Recently SMID-DB added reference spectra for different ascarosides, which are not commercially available [140].

3.1.1.3. Retention time (RT)

m/z values and tandem MS data are often not enough to identify metabolites. For example, the two ascarosides ascr#3 and oscr#3 have the same molecular formula and, therefore the same *m/z* and almost identical fragmentation spectra but very different retention times on a typical reversed-phase separation. RTs represent important orthogonal information for metabolite identification, which can be incorporated at various stages of the metabolite identification workflow. However, it usually is used at the end when comparing a reference standard against the measured metabolite. Knowledge about the employed chromatographic system enables to improve the metabolite identification at an early stage. Reversed-phase separation, for example, performs separation based on hydrophobicity, with polar metabolites eluting first and non-polar metabolites later. For example, this fact can be utilized using the octanol-water-partition coefficient for filtering annotations [319, 320].

Since metabolites cover a wide range of polarity, no single analytical method can cover the entire metabolome of a given sample or organism. Reversed-phase separation (RP) is used for the separation of hydrophobic substances. Two separation methods are commonly used:

The first one uses a gradient from water to organic solvents like acetonitrile (ACN) or methanol (MeOH) for the chromatographic separation of mid-hydrophobic metabolites, whereas the second uses gradients from water/ACN to 2-Propanol (iPrOH) and is typically employed for the separation of lipids.

The main driver for metabolite separation in RP is the partitioning between the hydrophobic stationary phase, e.g. octadecyl-modified silica particles and the hydrophilic mobile phase. Gradient elution toward solvents with higher elution strength (hydrophobicity, e.g. MeOH, ACN or iPrOH) also allows for eluting non-polar metabolites. Selectivity of separation can be fine-tuned by the addition of different functional groups or other ligands (e.g. phenyl-hexyl).

Analysis of hydrophilic metabolites can be performed using Hydrophilic Liquid Interaction Chromatography (HILIC). In contrast to RP, the separation mechanism of HILIC is not entirely understood. While the main driver is also the partitioning between two phases, the water-enriched hydrophilic stationary phase and the hydrophobic mobile phase, several secondary interactions also play important roles. These include ionic interaction, hydrogen bonds, and others. Therefore, the exact separation mechanism in HILIC is less well-defined and relies on the employed column and solvent. Metabolomics does not allow for a “one-size-fits-all” experimental protocol; hence, a diverse set of separation conditions are used in different laboratories. Additionally, comprehensive separation methods such as GCxGC and LCxLC are gaining more attention; they are mostly used by specialist laboratories [321, 322].

RTs are highly reproducible under identical chromatographic conditions within a single laboratory but can vary between different labs due to different factors like different equipment used or different separation conditions. Even when using nominally the same separation system, meaning the same column, solvent, and gradient, other LC systems yield different RTs due to differences in dwell volume, gradient delay volume, etc. Normalizing these differences would help boost metabolite identification by incorporating retention times in an early stage.

Recently, PredRet was introduced as a possibility to map RTs between different laboratories and chromatographic systems. This approach uses commonly identified metabolites between two systems to calculate a mapping function between them and use them to project retention times from one system to the other. Based on extensive cross-validation, the authors could show that this system helps to improve metabolite identification [323]. However, this system only works if several common metabolites are identified between two similar chromatographic methods. Projection between different LC-MS systems has been described by Abate-Pella *et al.*, using a test mixture to enable retention projection and back-calculation of retention times between multiple systems [324]. Different multi-segment gradients have been used, and different instrumental setups have been compared. Using the test mixture and retention projection enabled the accurate comparison between the participating laboratories. However, projection is only possible if commonly detected substances exist. Therefore, the prediction of retention times for substances that have not been measured before is of great interest. Different approaches for retention time prediction have been applied in metabolomics. Predicting RTs using Quantitative Structure-Retention Relationships (QSRR) can be seen as a particular case of Quantitative Structure-Property Relationships (QSPR) [325]. Predictions often focus on reversed phase (RP) separation, which is employed for the analysis of pharmaceutically active substances; here, retention times show a good correlation with the octanol/water partition coefficient logP. (Notably, logP values are usually only predicted via QSPR models, such as the XLOGP3 method [326]). HILIC was introduced by Alpert *et al.* and is gaining popularity due to its “orthogonal” analyte retention, compared to classical RP separations allowing the analysis of polar metabolites [327]. In contrast to RP, this correlation of logP with retention time in HILIC is only modest.

Early papers on QSRR date back to 1977; see [325, 328] for reviews. Since then, numerous articles have addressed the prediction of LC retention times for small molecules; Héberger [325] lists more than 100 papers from 1996 to 2006. Modern methods usually rely on supervised machine learning techniques such as linear regression, Support Vector Regression (SVR), random forests and regression trees, Neural Networks, partial least

squares regression, and feature selection for this purpose [329-334]. Usually, physicochemical parameters of the compounds are used as features, such as logP, charge, number of rotatable bonds, etc. Some methods are restricted to particular compound categories, such as nucleic acids, peptides, lipids, or polycyclic aromatic hydrocarbons [331, 335-341]. These methods reach much better prediction quality, as can be expected but are of limited use in general metabolomics. As noted, most papers focus on RP separation but fewer on HILIC separation [329, 342]. Randazzo *et al.* described a QSRR approach combining molecular descriptors with a Linear Solvation Energy Relationship model to predict retention times for steroid identification [343]. Three recent papers apply RT prediction for two-dimensional gas chromatography [344-346]. Most approaches use predicted retention times to filter candidate structures for metabolite identification and report results such as “half of the candidate structures were removed, and 95% of the correct identifications were retained”[347].

A significant problem of RT data, in contrast to MS data, is the only minimal and often incomplete collection of associated metadata. RT is influenced by many different parameters and is not an inherent property of the molecule but represents a property of the combination of the whole chromatographic system and the analytes of interest. Therefore, reporting of a single RT value is insufficient. Beside the analyte, the column with its exact stationary phase, dimensions, temperature, used solvents, flow rate, and gradient must be reported. Accurate reporting of the stationary phase is essential since several types of reversed-phase, and even C18 columns exist. Some of them have additional functional groups to enhance selectivity further. Reporting of chromatographic metadata needs to be improved, and standards have to be developed to make the use of RT in metabolite identification more widespread and reproducible.

3.1.1.4. Collisional Cross Section (CCS)

Another orthogonal parameter that can be utilized for metabolite and lipid identification is the Collisional Cross Section (CCS) measured by ion mobility. Ion mobility separation (IMS)

recently joined the metabolomics and lipidomics toolbox and different types of ion mobility instruments are available. IMS is a gas phase ion separation technique, often also referred to as gas-phase electrophoresis, separating ions by their differential travel through a drift cell filled with a buffer gas (typically N_2 or He) under the influence of an electrical field. Ions are separated based on their different mobility, which can be related to the molecules' CCS values. The use of drift tube ion mobility (DTIMS), traveling wave ion mobility (TWIMS), or trapped ion mobility (TIMS) allows determining a molecule's CCS. IM instruments can be classified according to their measurement principles. The DTIMS and TWIMS instruments are classified as time-dispersive, while TIMS is classified as field-dispersive (scanning), and DMA and FAIMS are spatially dispersive. CCS reflects the two-dimensional projection of the sphere that is formed by the randomly rotating ionized molecule in the gas phase. Ion mobility experiments can determine CCS values under low-field conditions by either calibrating or direct measurements using low-field uniform DTIMS.

In contrast to RT, CCS represents a molecular property. While the RT is dependent on the employed chromatographic system, including column, solvents, temperature, and others, CCS is more stable to changing instrumental conditions and can be compared within a specific range, even between different laboratories and instrumentations. CCSs are partially dependent on the *m/z* ratio of the measured molecule, but similar masses with different 3D structures can be resolved. At the current stage, it is not known how much exactly the CCS value adds to metabolite identification since databases are relatively small in comparison to MS^2 data. However, different *in silico* tools have been developed for the prediction with different accuracies. Due to the currently limited resolution of ion mobility instruments, it is mostly combined with chromatographic separation. In a theoretical investigation, Causon *et al.* calculated that at least a 1.5-1.8% difference in the CCS values is required for a resolution of 0.6, while baseline separation requires a minimum of 3.7-4.4% difference.

In drift tube IM, a uniform DC voltage gradient along the drift cell is used. The drift velocity V_d is proportional to the electric field E. In DTIMS, ions are separated in a static drift gas along

an electric field. The field is uniform with E/N values between 1 and 15 Td (1Td = 10^{-17} Vcm^2).

$$V_d = K \cdot E$$

The ion mobility can be calculated experimentally by measuring the ion's velocity or drift time, the time an ion requires to travel a drift cell with a known length L, and an applied field strength E. If the pressure in the drift cell, the drift time, and the voltage are known, the CCS can be calculated. Compared to drift times, CCS have the advantage that they are instrument independent and can be used as unique identifiers. The relation between the CCS value Ω and the drift time is determined by the Mason-Schamp equation

$$\Omega = \frac{(18\pi)^{1/2}}{16}\frac{ze}{(k_bT)^{1/2}}\left[\frac{1}{m_i}+\frac{1}{m_B}\right]^{1/2}\frac{t_dE}{L}\frac{760}{P}\frac{T}{273.2}\frac{1}{N}$$

In TWIMS, ions are separated by a sequence of symmetric potential waves which propagate through the drift cell. The E/N values are therefore varying with peal values between 50-160 Td. While in DTIMS and TWIMS, the drift gas is static, in TIMS, a pressure gradient is established in a differentially pumped area. Ions are held in place by an electric field counteracting the pressure difference. Ions are eluted sequentially by lowering the electrical field in a time-dependent manner. Typical E/N values are 45-85 Td.

CCS values can be used for metabolite identification in addition to MS and MS/MS and RT information and add an additional layer of security for annotations and identifications. It also allows for improving the analysis of unknown substances. If the CCS is plotted against *m/z,* typical trendlines for specific metabolite classes are observed, which can be used for enhanced annotation, even for lower-level annotations. In contrast to RT, CCS values can be predicted much easier. Different tools for using *ab initio* prediction utilizing different machine learning algorithms or quantum chemistry have been used [348-351].

3.1.1.5. NMR

In contrast to MS, NMR can be used for *de novo* structural elucidation. MS can give valuable information on potential building blocks present in a molecule, while NMR provides

information about their connectivity. However, NMR typically requires much more material compared to NMR. In the case of limited sample material, often the amount is not sufficient for structural elucidation or the material is not sufficiently pure. In this case, samples are typically pre-fractionated, e.g. using Flash or preparative chromatography. Since this contradicts the goal of systems biology and large-scale screening approaches, the use of NMR is not further discussed here. A possible solution is to fractionate reference samples and generate reference spectra of defined fractions and, in parallel, analyze them via LC-MS to achieve a cross-mapping and enable identification via NMR at a later stage. Since NMR is not covered in this thesis, it is not explained further.

3.1.2. Structural ambiguity

Although the combination of different methods can help to narrow down the number of potential metabolites, often a complete identification to a fully defined structure is not or only hardly possible. This is primarily true for stereochemistry. In cases of specific metabolites classes, such as flavonoids or lipids, only certain building blocks can be identified in typical non-targeted metabolomics approaches since the complete structure of small molecules cannot be fully established with current mass spectrometric methods, even with the highest resolution and precision and multistage MS experiments. Several types of ambiguities exist, which cannot be resolved solely based on MS, e.g. position of a double bond and its stereochemistry in lipids or the exact position of hydroxyl groups in flavonoids. Often this requires specialist methods, not available at every metabolomics/lipidomics laboratory, e.g. Ozone-induced dissocation (OzID) [352]. Missing full identification leads to a problem in reporting metabolites and lipids, e.g. using standard data exchange formats such as mzTab [353]. This format requires an identifier from a common compound database, e.g. HMDB, ChEBI, PubChem, LipidHome, or LIPID MAPS, and optionally a chemical formula, SMILES or InChI. Structural databases, like most of the above mentioned, allow only entire molecule structures to be deposited. ChEBI goes beyond this by allowing the deposition of general group structures or more general structures [202]. Ideally, structures should be unambiguously resolvable down to the complete structural level. However, in metabolomics

and lipidomics, it is still common to report full structural details, although no experimental evidence was collected, which leads to an overestimation of the identification and following biological interpretation. Different schemes for reporting the level of identification have been proposed, and even numerical systems have been suggested (see above). Typically, the highest identification level is achieved by comparing the compound of interest and a reference standard based on two orthogonal and independent physicochemical properties. But often, this still leaves open different stereochemical isomers (e.g. D- and L-Tryptophan) since they will have identical retention times and fragmentation patterns in standard achiral separation conditions used in metabolomics and lipidomics.

For lipids specifically, ambiguities exist concerning the position of fatty acid side chains, the position of double bonds inside chains, their stereochemistry, and how to encode them in a consistent naming scheme. The current terminology uses lipid category abbreviations (e.g. PC) and different levels of structural details [205, 207, 247]. This nomenclature and classification scheme groups lipids into defined categories (e.g. Glycerophospholipids), main lipid classes (e.g. PC), subclass (Diacylglycerophospholipid), species, fatty acids scan species, sub-species and isomer level. The most detailed level (isomer) identifies the exact position of side chains, double bonds, and stereochemistry, which cannot be achieved for high-throughput applications with current analytical methods. Therefore, a more functional and MS-friendly nomenclature has been invented, dividing lipids according to their class and acyl side chains. The nomenclature considers the structural detail that specific experiments can capture. Often only the lipid class and its sum composition can be deduced from MS^1 and MS^2 experiments. The identified lipid would be reported accordingly, e.g. for a triacylglycerol with a total of 52 Carbon atoms and one double bond; the name would be reported as TG 52:1. If MS^2 analysis was able to reveal fatty acid bound to the glycerol backbone, the identification can be reported as TG 16:0_18:0_18:1 if is known that one of the fatty acid chain consists of sixteen carbon atoms, the second fatty acid chain of eighteen, and the third one of eighteen carbon atoms with one double-bond at an arbitrary position. More detailed experiments finally resolve the position of the fatty acids and the position of the

double bond (e.g. MS^3, etc.…) [354]. However, LipidMaps only contains full structures; ChEBI links different chemical entities in its database via an ontology system. Using this system, chemically related compounds are linked via functional parents. A similar system is used by LipidMAPS, which groups lipids according to their lipid classes. However, both systems are missing important intermediate steps in their hierarchy or lack structural representation. For example, ChEBI contains an entry for the lipid PC(36:2) (ChEBI:64433) and links PC(18:0/18:2) (ChEBI:136063) and other more detailed structures to it, but has only generic SMILES for diacyl phosphatidylcholine associated with it, while molecular formula and mass are matching. Furthermore, different species like PC(16:0/20:2) are missing, which make up the same sum composition. This discrepancy must be resolved by allowing extended SMILES to be deposited in databases. The so far best hierarchical representation is provided by the SwissLipids database [289]. Ontologies like these can be used for enhanced mapping of lipid structures onto metabolic pathways [204].

In principle, the same issues arise not only in lipidomics but similarly for other small molecules (<1kDa) in environmental chemistry, glycomics, natural product chemistry (flavonoids and terpenes), and metabolomics. In these fields, it can be referred to unknown stereochemistry, the unknown position of functional groups, unknown sugar moieties, etc. While a neutral loss in a fragmentation spectrum of 162.0528 Da is often indicated for a hexose group position in case of multiple possible sites, and the identity of the hexose cannot be revealed. One prominent example of such structural ambiguity in metabolite identification is flavonoids and their glycosides. Using MS/MS, often only the number, but not the specific position of hydroxyl groups attached to different rings of the flavonoid structure or the number of glycosylations but not their position nor their identity can be determined. One example would be quercetin (ChEBI:16243) and morin (ChEBI:75092). Both are classified as pentahydroxy flavones (CHEBI:25883), a rather general description for identification purposes. A typical example for *C. elegans* would be Δ^4- (CHEBI:78686) and Δ^7-dafachronic acid (CHEBI:78699), which cannot be separated with generic LC-MS methods as used in non-targeted metabolomics nor can they be separated by different fragmentation.

Further examples are oscr and ascr molecules, which show a similar fragmentation pattern but have a difference in their linkage of the fatty acid tail to the ascarylose core (either ω- or ω-1-). One future possibility would be to report, for example, Markush structures instead of complete structures [355]. Such structures are typically used in patent applications to cover all possible structures in a single depiction instead of drawing each structure individually. Multiple structures can be collapsed into a single Markush structure. However, no open-source chemoinformatics tools exist at the moment; depending on the input structures, results might not be human-readable anymore.

3.1.3. The need for metabolites identification for systems biology

Systems biology approaches require metabolite identification. In most cases the analysis is performed on the pathway level, and combined mapping of transcriptomics, proteomics, and metabolomics/lipidomics on pathways will be performed, and results will be analyzed in combination. Without correct metabolite identification, such mapping cannot be performed. Though ambiguities might still exist, ontology-based approaches allow mapping even with only partial structures and to judge the goodness of mapping [204]. Besides pathway mapping, correlation analysis, and other tools are important for the analysis of large-scale systems biology studies. Although correlation values can be established between two completely unknown molecules, a biological relation can only be established after the identification of at least one of the partners. Likewise, molecular networks such as those generated by GNPS can be used [317, 356]. In such networks, metabolite identities can be used to enhance the identification of other non-identified nodes [318, 357].

Lastly, metabolite identifications allow combining different datasets based on the metabolites, which is not or only partially possible based on mass spectrometric raw data. This will become more and more important as more studies on *C. elegans* will become publicly available. Integration of different studies increases the power of statistical analysis and the information content but currently suffers from the missing standardization of metabolomics

methods. Standardized metabolite identification and reporting represent one of the first and most important steps toward this integration.

3.2. Retention time indexing

Sharing of RTs for cross-laboratory comparison would be helpful for enhanced metabolite identification. However, the lack of standardization of chromatographic conditions and different analytical instrumentation complicate the (re-)use of retention information. In contrast to *m/z* or CCS, which represent molecular properties, RTs are system properties that arise from the combination of chromatographic equipment, employed column and solvent, and separation conditions (e.g. flow rate, temperature, etc.). Even nominally the same separation conditions on two different instruments can yield differences in RTs due to experimental factors like dead volume, gradient delay volume, etc. Furthermore, even within a single lab RTs can shift substantially due to column and solvent batches (differences in solvent composition, pH adjustment etc.) and the degree of deterioration of the chromatographic column. Therefore, retention information is not used regularly for cross-laboratory metabolite identification. Different approaches to tackle this problem have been developed.

In GC RTs are normalized to retention indices (RI) by referencing the retention time of a given substance to a set of reference substances. This retention time indexing (RTI) is well established and allows cross-referencing different separations performed under similar but not the same conditions. The Kovats index is the most used RTI system in GC and has also been applied in metabolomics and uses a homologous series of n-alkanes spiked to the sample as references [358, 359].

Different RTI systems have been proposed for LC-based separation, each providing distinct advantages and disadvantages. Aderjan & Bogusz introduced an RTI system based on a series of 1-nitroalkanes [360]. This substance class shows strong absorbance between 200 and 230 nm. However, they ionize only in negative ionization mode in electrospray MS. Different other substances have been suggested as RTI markers, e.g. alkyl phenones, phenolic esters, and others reviewed elsewhere [361]. Nitroalkanes and fatty acid amides have also been used for RTI in metabolomics [362]. These RTI systems have also been

combined with different *in silico* analysis methods for tandem MS spectra, CFM-ID, CSI:FingerID, MassFrontier, and MetFrag [313, 363, 364]. Based on a set of measured RI values, an artificial neural network was trained to predict RI values for candidates from *in silico* methods. This combined method was able to improve the average rank of candidate molecular structures [365]. Recently, Zheng *et al.* established an RTI system for derivatized molecules. 2-dimethylamino ethylamine (DMED) labeled fatty acids served as indexing substances for DMED-labeled carboxylic acids [366]. Amine compounds were labeled with 4-(*N,N*-dimethylamino)phenyl isothiocyanate (DMAP), and DMAP-labeled fatty amines served as RI standards. Based on their RTI system, the authors compared different chromatographic setups and could show that RI is much more comparable then RT. However, most metabolomic experiments detect metabolites in their native, underivatized state. Therefore, a solution for RTI for unlabeled substances is required. A further disadvantage of the presented substances for RTI is that for positive and negative ionization modes, different substances are used, resulting in two different sets of RI databases. *N*-alkyl pyridinium 3-sulfonates (NAPS) have been suggested as promising candidates for the normalization of RT data by conversion to RI. Recently, they have been used for the normalization of LC-MS mycotoxin determination [367]. The properties of the NAPS make them an ideal candidate for testing as an RTI system for the normalization of retention time information in LC-MS-based metabolomics. In different studies, the suitability for normalization within single laboratory or multiple laboratories was examined.

3.2.1. Material and Methods

3.2.1.1. Chemicals

Acetonitrile (ACN), methanol (MeOH), formic acid, and chloroform ($CHCl_3$) were purchased from Sigma-Aldrich and were of LC-MS grade (Sigma-Aldrich, Taufkirchen, Germany). Metabolite standards used in this study were derived from the Mass Spectrometry Metabolite Library of Standards (MSMLS) (Sigma-Aldrich, Taufkirchen, Germany). Standards were prepared as indicated in the MSMLS manual. Metabolites are summarized in SI Table 1. Additional solutions from individual metabolites were prepared as 1 mg/mL stock solutions in

suitable solvents. Additional metabolites are summarized in SI Table 2. The NAPS RI standards in the form of a reference material (RM-RILC) composed of a mixture of 20 NAPS in solution (100 µM each) were provided by the National Research Council Canada (Halifax, NS, Canada, (https://www.nrc-cnrc.gc.ca/eng/solutions/advisory/crm/list_product.html). This mixture is referred to as NAPS in the following paragraphs. Different chromatographic columns have been used in different studies. Table 14 summarizes the information on the employed columns and the respective studies they were used in.

Table 14: Chromatographic columns used in the different studies related to retention time indexing.

Column	Study
Supelco Ascentis Express C18 column (100 mm x 2.1 mm, 2.0 µm)	Intra-lab pilot
Supelco Acsenctis Express C18 column (150 mm x 2.1 mm, 3.0 µm)	Intra-lab pilot
Waters Acquity BEH C18, 100 mm x 2.1 mm, 1.7 µm (Batch 1)	Cross-column pilot Interlab pilot
Waters Acquity BEH C18, 100 mm x 2.1 mm, 1.7 µm (Batch 2)	Cross-column pilot
Waters Acquity BEH Shield RP 18, 100 mm x 2.1 mm, 1.7 µm	Cross-column pilot
Waters Acquity BEH CSH C18, 100 mm x 2.1 mm, 1.7 µm	Cross-column pilot
Waters Acquity Cortecs C18, 100 mm x 2.1 mm, 1.6 µm	Cross-column pilot
Waters Acquity Cortecs C18+, 100 mm x 2.1 mm, 1.6 µm	Cross-column pilot
Phenomenex Kintex C18, 150 mm x 2.1 mm 1.7 µm	Interlab full

3.2.1.2. Intra-laboratory pilot study

For the first intra-laboratory pilot study, three different LC-MS setups were used. The first setup consisted of a Waters Acquity UPLC (Waters, Eschborn, Germany) coupled to Bruker maXis UHR-ToF-MS (Bruker Daltonics, Bremen, Germany). Metabolite standards were separated on a Supelco Ascentis Express C18 column (100 mm x 2.1 mm, 2.0 µm) (Sigma-Aldrich, Taufkirchen, Germany). Eluent A consisted of 100% H_2O + 0.1% formic acid and eluent B of 100% ACN + 0.1% formic acid. The flow rate was set to 0.3 ml/min, and column

temperature was maintained at 40°C. After 2 minutes of 5% B, %B was increased linearly to 99.9% within 15 min and held for 3 minutes. After returning to the initial conditions, the column was re-equilibrated for 3 minutes. 5µL of the sample were injected via partial loop injection. Different methods for systematic evaluation of flow rate and temperature influences were derived from this standard method. Metabolite standards were detected in positive and negative ionization mode with data-dependent acquisition (DDA) of tandem MS. Source parameters were as follows: End Plate Offset = 500 V, Capillary = 4500 V, Nebulizer = 2.0 bar, Dry Gas = 10.0 L/min, Dry Temp = 200°C, Mass Range = 50 – 1500. The second LC-MS setup was an Agilent 1200 HPLC equipped with a quaternary pump (Agilent Technologies, Waldbronn, Germany) coupled to a Bruker maXis UHR-ToF-MS (Bruker Daltonics, Bremen, Germany). Separation of metabolites standards was performed on a Supelco Acsenctis Express C18 column (150 mm x 2.1 mm, 3.0 µm) (Sigma-Aldrich, Taufkirchen, Germany) using the same eluents as above. Flowrate was set to 0.4 mL/min. After 4.5 minutes of 5% A, %B was increased linearly to 99.9% within 35.5 min and held for 6 minutes. MS parameters were the same as for the first system. The third LC-MS setup used the same column and gradient as the second but performed the separation on the hardware from the first setup (Waters Acquity UPLC), differing in the gradient formation and delay volume. MS parameters were the same as for the first system.

Fifty-two mixtures of up to eight non-isomeric or -isobaric metabolite standards were prepared from the MSMLS and measured for the initial database construction. NAPS were injected before and after samples of one plate. For all other experiments, mixtures from one plate were injected. Peaks were manually picked in Data Analysis 4.4 (Bruker Daltonics, Bremen, Germany) by creating extracted ion chromatograms. MS and MS/MS spectra were used for verification. All further analyses and calculations were performed in Microsoft Excel 365 and R 4.2.1 [368].

As biological samples, extracts from *C. elegans* and mouse plasma were used. Mixed-stage *C. elegans* were grown in a liquid culture fed with *E. coli* NA22 and harvested by

centrifugation. Metabolites were extracted with 50% MeOH, according to Witting *et al.* [185]. After the extraction solvent was evaporated, samples were redissolved in 20% ACN to achieve an estimated concentration of ~10.000 worms/mL. Proteins from mouse plasma were precipitated by mixing 500 µL plasma with 1500 µL ice-cold ACN. Samples were vigorously vortexed and centrifuged for 15 min at 13,000 rpm at 4°C. The supernatant was transferred to a fresh reaction tube, and the solvent was evaporated. The sample was redissolved in 500 µL 20% ACN. Aliquots of *C. elegans* metabolite and mouse plasma extracts were spiked with NAPS at levels of 1:20, 1:40, and 1:80.

Data were processed with Genedata Expressionist for MS 13.5, which included *m/z* recalibration, noise reduction, RT alignment, peak picking, and isotope grouping. Results were exported as a .xlsx file and further processed in R as described below. Annotation of metabolites was performed with the `MetaboAnnotation` package (https://github.com/rformassspectrometry/MetaboAnnotation) [369].

3.2.2. Cross-Column pilot study

In this study, a Waters Acquity UPLC (Waters, Eschborn, Germany) was coupled to a Bruker maXis UHR-ToF-MS (Bruker Daltonics, Bremen, Germany). The MS parameters were the same as above. Eluents were 100% H_2O + 0.1% formic acid and 100% ACN + 0.1% formic acid. Three different gradients were used, as summarized in Table 15.

Table 15: Gradients used in the Cross-column pilot study. Individual steps of %B were kept and only times were varied.

	Time (min)		
%B	**Gradient 1**	**Gradient 2**	**Gradient 3**
5	0	0	0
5	1.12	1.12	1.12
99.5	6.4	11.7	17
99.5	10	15.3	20.6

Columns listed in Table 14 and referred to be used in the cross-column study were used. Standard solutions from 99 metabolites (SI Table 2) were pooled into five mixes which do not contain any isomeric or isobaric substances. Peaks were manually picked in Data Analysis 4.4 (Bruker Daltonics) by creating extracted ion chromatograms. MS and MS/MS spectra were used for verification. All further calculations were performed in R 4.2.1 [368].

3.2.2.1. Interlaboratory pilot study

In this pilot study, five different LC-MS systems in five other laboratories were used. The first laboratory used a Waters Acquity UPLC (Waters, Eschborn, Germany) coupled to a Bruker maXis UHR-ToF-MS (Bruker Daltonics, Bremen, Germany). The second lab utilized a Waters Acquity H-Class UPLC (Waters, Milford, MA, USA) coupled to Bruker maXis 3G UHR-ToF-MS (Bruker Daltonics, Bremen, Germany). The third lab used a Waters Acquity H-Class UPLC (Waters, Milford, MA, USA) coupled to a Waters Xevo TQ (Water, Manchester, UK), while the fourth lab employed an Agilent 1290 UHPLC (Agilent Technologies, Waldbronn, Germany) coupled to a Leco Citius HR-ToF (Leco, St. Joseph, MI, USA). The last laboratory used a Thermo Scientific Dionex Ultimate 3000 (Thermo Scientific, Germering, Germany) coupled to a Bruker impact II Q-TOF-MS (Bruker Daltonics, Bremen, Germany).

Retention times of the same 99 metabolite standards as in the cross-column comparison were determined using a Waters Acquity BEH C18 column (100 mm x 2.1 mm, 1.7µm) and a water-ACN gradient with eluent A begin 100% H_2O + 0.1% formic acid and B 100% ACN + 0.1% formic acid. Gradient conditions were as follows, after an initial isocratic phase of 5% B for 1.1 minutes, %B was increased linearly to 99.5% B in 5.3 minutes and held for 4.6 minutes. Each participating lab received prepared stock solutions made from metabolite standards at a concentration of 1 mg/mL in order to tune MS parameters or MRM transitions individually. NAPS were supplied as a ready-to-use reference mixture.

3.2.2.2. Interlaboratory study

The interlaboratory study was performed between 3 laboratories running 4 different LC-MS systems. The first lab conducted experiments using an Agilent 1290 Infinity II UHPLC

coupled to an Agilent 6560 IM-QTOF-MS (Agilent Technologies, Waldbronn, Germany) or a Waters Acquity UPLC (Waters, Eschborn, Germany) coupled to a Bruker maXis UHR-ToF-MS (Bruker Daltonics, Bremen, Germany). The second laboratory used a Waters H-Class Acquity UPLC coupled to a Waters Vion TWIMS-QToF-MS (Waters, Manchester, UK). The third laboratory used a Waters Acquity UPLC (Waters, Eschborn, Germany) coupled to a Thermo LTQ XL ion trap MS (Thermo Scientific, Bremen, Germany).

In all laboratories, a Phenomenex Kinetex C18 column (150 mm x 2.1 mm ID, 1.7 µm particle size) (Phenomenex, Aschaffenburg, Germany) with a fitting SecurityGuard Ultra was used. Eluent A consisted of 100% H_2O + 0.1 % formic acid and eluent B of 100 % ACN + 0.1% formic acid. The following gradient was used: 2% B at 0 min, 100% B at 14 min, 100% B at 17 min, 2% B at 17.1 min, and 7.9 minutes column re-equilibration. The column temperature was set to 30°C and the flow rate to 0.3 mL/min. MS detection was performed in positive and negative ionization modes.

Metabolites from the MSMLS library (Sigma-Aldrich, Taufkirchen, Germany) were used. Each laboratory was supplied with a pool from each plate together with NAPS standard solution. After some initial blank injections, NAPS was injected, followed by the seven plate pools. This was repeated five times and one final NAPS injection, so every plate injection was bracketed by two NAPS injections.

Peaks were picked individually in each laboratory, and RT data were collected in a central Excel sheet for further processing. In the first laboratory, additional injections of a *C. elegans* extract prepared as mentioned above were performed. Data from these experiments, together with NAPS injection, were processed using Genedata Expressionist for MS 13.5 as described above. Further handling of all data was performed in R 4.2.1 [368].

3.2.2.3. Retention time indexing

A function for the conversion of RT to RI has been implemented in the `MetaboCoreUtils` package (https://github.com/rformassspectrometry/MetaboCoreUtils) [369]. This function accepts a `vector` of retention times, a `data.frame` with two columns, one corresponding

to the RT and the other to the RI of reference substances, and a conversion function. The default function uses linear interpolation between the two bracketing RI standards according to the following equation:

$$RI = RI_0 + (RI_1 - RI_0)\frac{RT - RT_0}{RT_1 - RT_0}$$

RI_0 and RI_1 denominate the RI of the bracketing NAPS, which is the number of carbons in the alkyl chain multiplied by 100, and RT_0 and RT_1 are the corresponding retention time. RT is the retention time of the substance for which the RI shall be calculated. However, custom functions can be supplied, e.g. using spline functions. A `vector` of the same length as the input RTs with the respective RI values is returned.

Two alternative functions for indexing have been implemented. The first one uses `spline()` for interpolation, while the second is based on the `aspline()` function from the `Akima` package [370].

3.2.3. Results and Discussion

3.2.3.1. Intra-laboratory pilot study

NAPS have been proposed by Quilliam *et al.* as applicable reference standards for indexing in reversed-phase LC-MS [367, 371]. Their properties and easy access due to simple synthetic procedures make them ideal candidates for this purpose. Synthesis and use of NAPS have been patented, and they can only be obtained from the NRC Canada. Synthetically they can be produced from 3-pyridine sulfonic acid reacting with N-alkyl halides. Due to the two permanent and oppositely charged groups (quaternary, aromatic imine, and sulfonate), the retention of NAPS is virtually independent of the separation pH. Furthermore, they can be detected in positive and negative ionization modes and UV detectors due to the aromatic ring. In the employed chromatographic setups, NAPS ionize as protonated $[M+H]^+$ ions in positive ionization mode and as $[M-H]^-$ and $[M+HCOO]^-$ adduct ions in negative ionization mode. Collision-induced fragmentation yields a common fragment of m/z 160.0063 in positive mode ($[C_5H_6NO_3S]^+$) and m/z 79.9579 in negative mode ($[SO_3]^-$), which makes the

substances also useful for MRM or DIA-MS/MS workflows. Typical fragmentation patterns of C10-NAPS are depicted in Figure 10. At higher concentrations also, gas-phase multimer formation (e.g. $[2M+H]^+$) as well as sodium adducts $[M+Na]^+$ and sodiated multimers (e.g. $[2M+Na]^+$) can be observed. The employed NAPS mixture (RM-RILC) consists of 20 homologs with a length of the *N*-alkyl chain from 1 to 20. The retention indices of the standards are denoted as 100 times the number of carbons in the alkyl chain (i.e., RI = 100 to 2000). The first three standards (C1-3, RI 100-300) elute in or close to the void volume, and RTs of the C1-C3 analogs were nearly identical under all separation conditions studied in this and the following sections. Therefore, robust RIs could only be calculated for substances for which RTs were higher than the RT of the C3-NAPS standard.

Initial experiments were conducted on a setup equipped with a high-pressure binary gradient pump capable of pumping at a maximum back-pressure of 1034 bar (LC-MS System 1). An initial RT/RI library was constructed by measuring the metabolite standards contained in the commercial MSMLS library. This library is sold in 96-well plates, and to reduce the number of injections, the rows of each of the seven 96-well plates were pooled to yield mixtures of standard not containing any isomeric or isobaric metabolites. The injections from one 96-well plate (e.g. Plate 1 A-H) were bracketed by one injection of NAPS before and after each block. From a total of 619 metabolite standards, 490 could be detected with the employed settings, 313 in both ion modes, 153 only in positive ion mode, and 24 only in negative ion mode. RTs were highly reproducible with a maximum of 6.1% and 5.4% relative standard deviation in positive and negative modes, respectively. Filtering for metabolites only eluting after C3-NAPS, a total of 219 metabolites remained for further analysis.

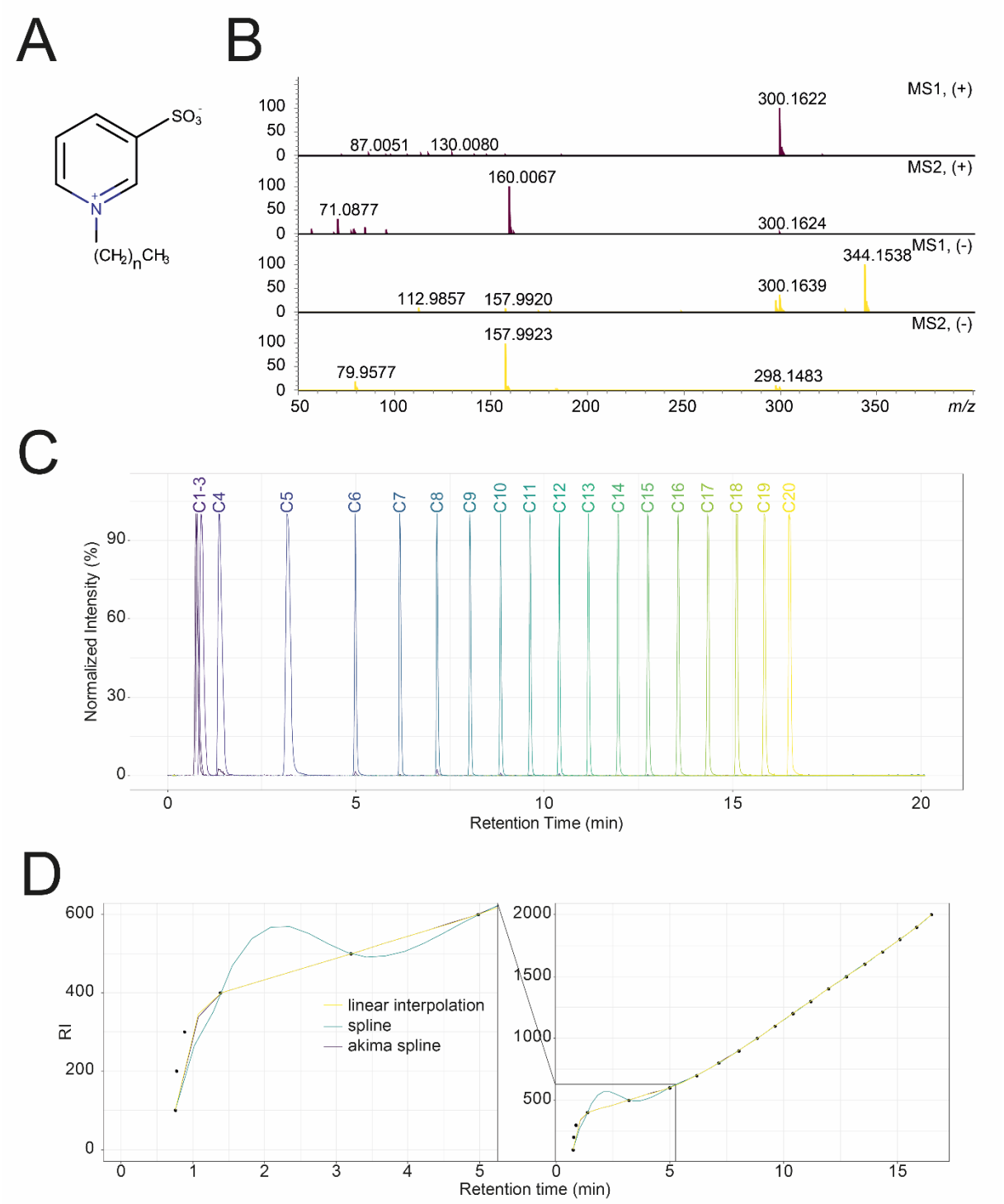


Figure 10: (A) General structure of the NAPS standards; (B) Positive and negative MS and MS/MS spectra of C10 NAPS. Fragments m/z 160.0067 in positive and 79.9579 are common to all NAPS. In case of negative ionization the $[M+HCOO]^-$ was fragmented. (C) Elution pattern of NAPS C1-C20 on the employed C18 conditions. NAPS C1-C3 elute in the void volume. (D) Different fitting functions can be used to convert RTs to RI. For higher RT linear interpolation, cubic splines or Akima cubic splines lead to almost identical results. However, for lower RTs, cubic splines can overshoot due to strong changes in the gradient due to the requirement of steadiness at each nodal point. This is specific to the employed analysis conditions. Use of RP columns compatible with 100% aqueous eluents might increase retention of C1- to C3-NAPS and early eluting analytes, which should improve fitting

Next, different possibilities for the conversion of RT to RIs were examined. The above-described functions from the `MetaboCoreUtils` package have been employed. For conversion to RI, either the NAPS injection before, after, or the average RT of both was used together with linear interpolation, spline, or Akima splines interpolation. Each of the three performed replicate injections was converted individually, and the mean, standard deviation, and relative standard deviation were calculated from the three replicates. Using either the injections of NAPS before, after, or their average had no significant influence, and the difference was in the range of the standard deviations, ranging between 1 and 6 RI units. Throughout all the following work, only the average of the bracketing NAPS injections was used. Subsequently, the different interpolation functions were compared. Akima splines were selected since they were suggested by Renaud et al. [367]. The comparison showed that, especially for RI values <750, systematic differences could be observed. Spline interpolation tends to overshoot in the region from the C4-NAPS to C6-NAPS due to big changes in the gradient and the requirement of steadiness at each nodal point. This requirement is dropped in the case of Akima splines, leading to a lower overshooting and better agreement with linear interpolation. In order to also make use of metabolite in the range between the C3- to C6-NAPS, linear interpolation was used throughout all remaining work and studies.

After the construction of an initial RT/RI library, it was first checked if the normalization of RTs to RIs can compensate for differences in experimental settings such as flow rate. Based on the method for the construction of the initial RT/RI library, the flow rate was systematically varied. Four different additional flow rates were tested (0.20, 0.25, 0.35, and 0.40 mL/min) on the same column with otherwise the same settings. Changes in flow rate do not interfere with the elution order, which is known from the previous experiments. Therefore, a pooled sample per plate was used. As suspected, a systematic shift of RTs based on the flow rate was observed, with lower flow rates showing a shift towards higher RTs, while higher flow rates resulted in faster elution. Differences ranged from +50% to -30%, which makes the use of RT for cross-separation system identification not possible. Conversion of RT to RI was able to normalize the data, and deviations were reduced to ≤ 5%. Eighteen substances were

excluded since they showed variations higher than 5%, and investigating their raw data showed that they were of very low abundance with low-quality MS^1 and MS^2 data, which potentially led to the wrong determination of their RT. Plotting the RIs of the different conditions against each other, slopes close to 1 and $R^2 > 0.9$ were obtained.

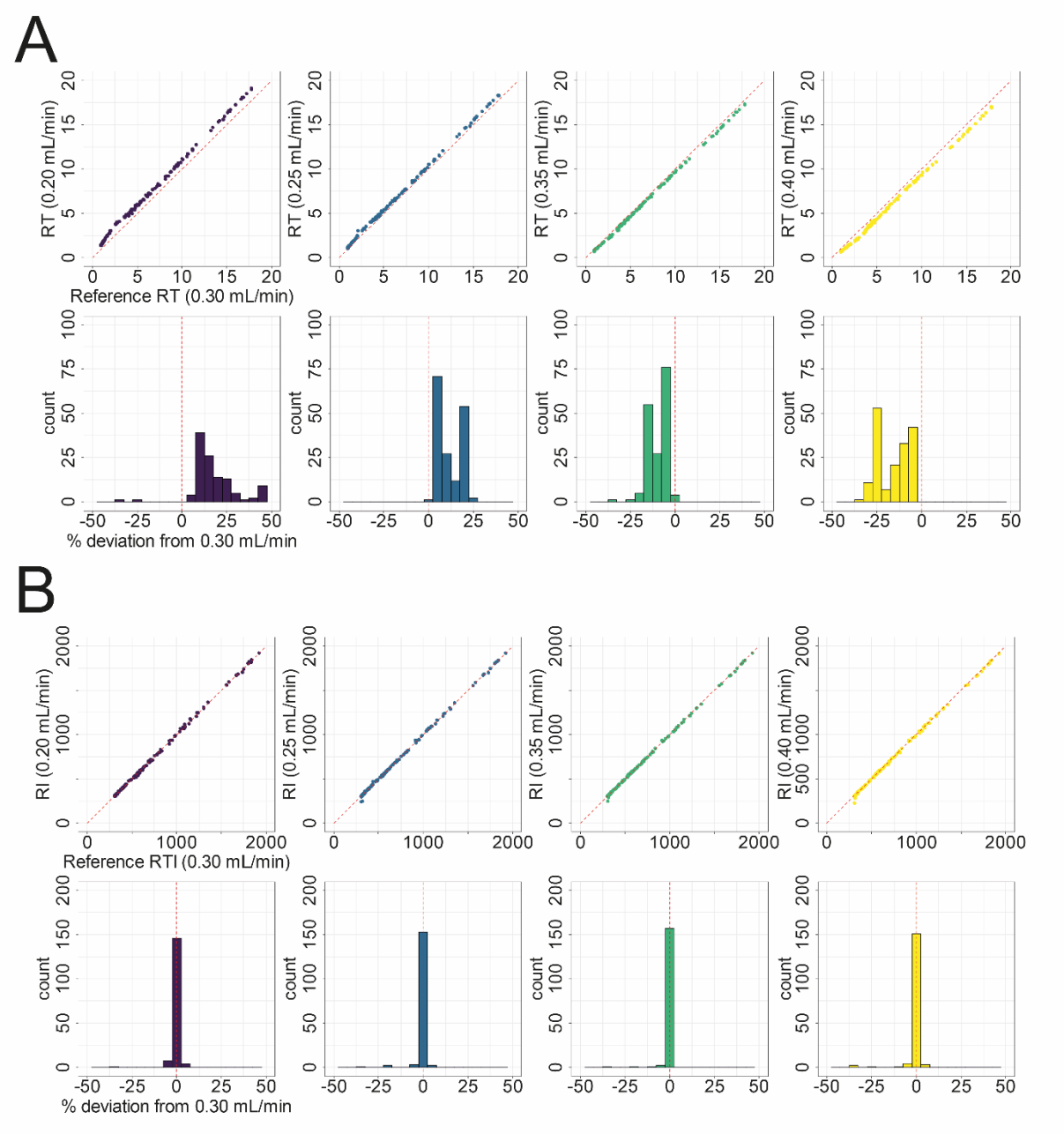


Figure 11: Deviation of RT under different flowrates compared to standard condition (0.30 mL/min). The upper row shows the scatter plot of the RTs from each different flow rate plotted against the reference condition. The red dashed line indicates the diagonal, which represents a perfect fit. The lower row show histograms of the deviation in %. (B) Deviation of RI under different flowrates. Plots are similar to (A), except that RI is used instead of RT. All plots indicate that the conversion of RT to RI enables the normalization of the different flow rates.

A further experiment was conducted to check how far RI allows the comparison of data obtained on different types of LC setups. As a second system, a low-pressure gradient quaternary pump with a maximum pressure of 400 bar was used. The same column chemistry and eluents were used, and similar to the previous experiment, standards were injected as pools for each plate. As anticipated a strong difference between the UHPLC and HPLC systems was observed for the RT, but a good agreement for the RI was observed. The slope of the plot of the different RIs was 1.004, and the R^2 was 0.998, with 74% of all detected standards shared between the two setups having a deviation between -5.0 and +5.0%.

So far, only mixtures of chemical standards have been analyzed. In the next step, the performance of the NAPS RTI system in combination with biological samples was tested. The most accurate application of NAPS would be to spike or co-inject them with all samples individually; however, due to the two strong permanent charges of the NAPS they can cause ion suppression. To test to which extent NAPS lead to ion suppression and to generate an independent data set for the evaluation of the RI, the HPLC format column was used in the UHPLC system. This setup has different characteristics compared to the previous two, and obtained RTs were different. For the following evaluations, *C. elegans* and mouse plasma metabolite extract have been used as biological matrices. NAPS were diluted with the two matrices or with 20% ACN at factors of 1:20, 1:40, and 1:80 and injected into the LC-MS. No significant shift in RTs of the NAPS between injections in pure solvent or in the matrix was observed.

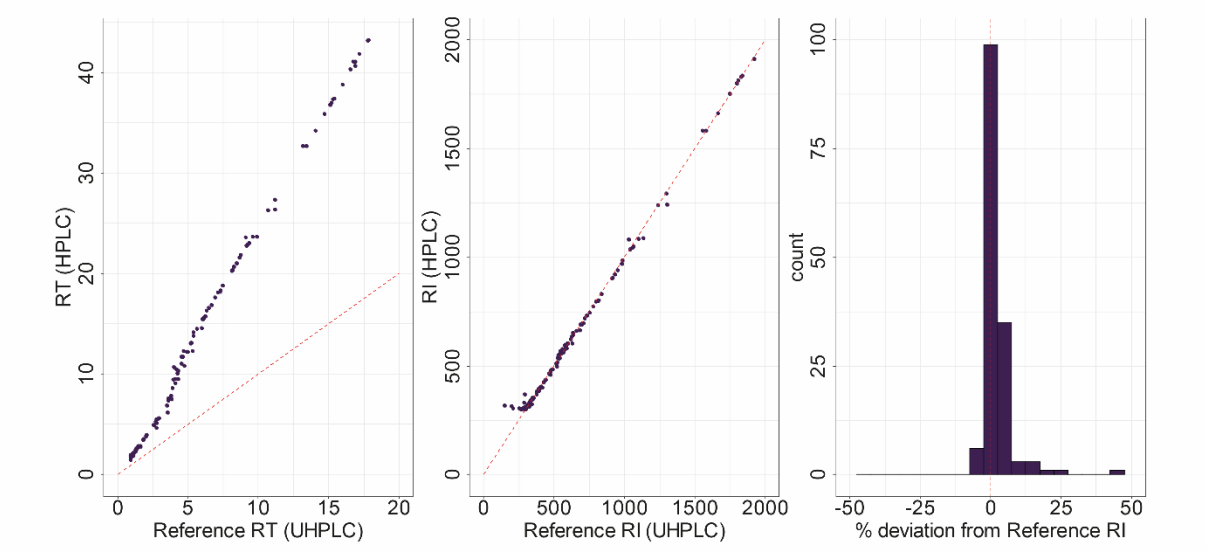


Figure 12: Comparison of HPLC and UHPLC on the basis of RT shows the expected deviation, while normalization using NAPS RTI made data comparable between the different column formats with most deviations between -5% and +5%.

First, the ion species detected in biological samples were evaluated. Following adducts were included: $[M+H]^+$, $[2M+H]^+$, $[3M+H]^+$, $[M+Na]^+$, $[2M+Na]^+$ and $[3M+Na]^+$ in positive ion mode and $[M-H]^-$, $[2M-H]^-$, $[3M-H]^-$, $[M+HCOO]^-$, $[2M+HCOO]^-$ and $[3M+HCOO]^-$ in negative ion mode. At a dilution of 1:20, most of the tested adducts in positive and negative ion mode could be detected, with the exception of NAPS 100 to 300, for which only the $[M+H]^+$ or $[M+HCOO]^-$ could be detected. $[3M-H]^-$ adducts were not detected, and $[2M-H]^-$ only for middle and long chains (C8-20).

Second, potential ion suppression was evaluated by comparing the intensity of metabolite features of unspiked with the spiked matrix. The first three NAPS standards (100-300) are coeluting together with many polar metabolites in or near the void volume, where high suppression is usually observed. Therefore, only NAPS standards with an RI > 300 were evaluated. Peaks eluting in the range of ± 0.20 min around the RT of the respective NAPS were evaluated, and their maximum intensities were compared against the non-spiked matrix. Internal *m/z* recalibration, chromatographic alignment, peak peaking, and isotope clustering have been performed in Genedata Expressionist for MS 13.5, and the maximum intensity for each isotope cluster was exported. Relative values compared to the non-spiked matrix were calculated for each cluster, and 100% would indicate no ion suppression, values

<100% ion suppression, and values >100% ion enhancement. NAPS chromatographic peaks generally span a width of about 0.2 minutes. To investigate the effect of ion suppression, the RT distance of a feature to the RT of the respective NAPS was plotted. With increasing, dilution factor ion suppression is reduced, as shown in Figure 4 for the C15 NAPS in positive ionization mode. For both investigated matrices, higher dilutions lead to reduced ion suppression, although 100% is never fully achieved (Figure 13A). Similar trends were seen in negative ionization mode (data not shown). However, ion suppression effects must be evaluated carefully for each matrix and LC-MS setup. If suppression effects might still be too strong, a separate batch of QC samples could be spiked with NAPS to determine RTs in matrix but not directly affect the biological and QC samples of the metabolomics studies. This spiked QC could be injected after each QC sample every ten samples.

Finally, as part of this intra-laboratory study the usability of RIs for annotation of metabolites was tested. The data obtained from the biological matrices was used as test case. Compared to the measurements on the second LC-MS setup NAPS showed RT differences of up to 3 min, which is due to reduced dead volume and differences in the formation of the gradient (high-pressure binary gradient vs. low-pressure gradient quaternary pump). Annotation of features detected in *C. elegans* was performed using the `matchMz()` function from the `MetaboAnnotation` package using *m/z* alone, or *m/z* and RI combined. The database for annotation was either the metabolites detected on the first (UHPLC) or the second setup (HPLC).

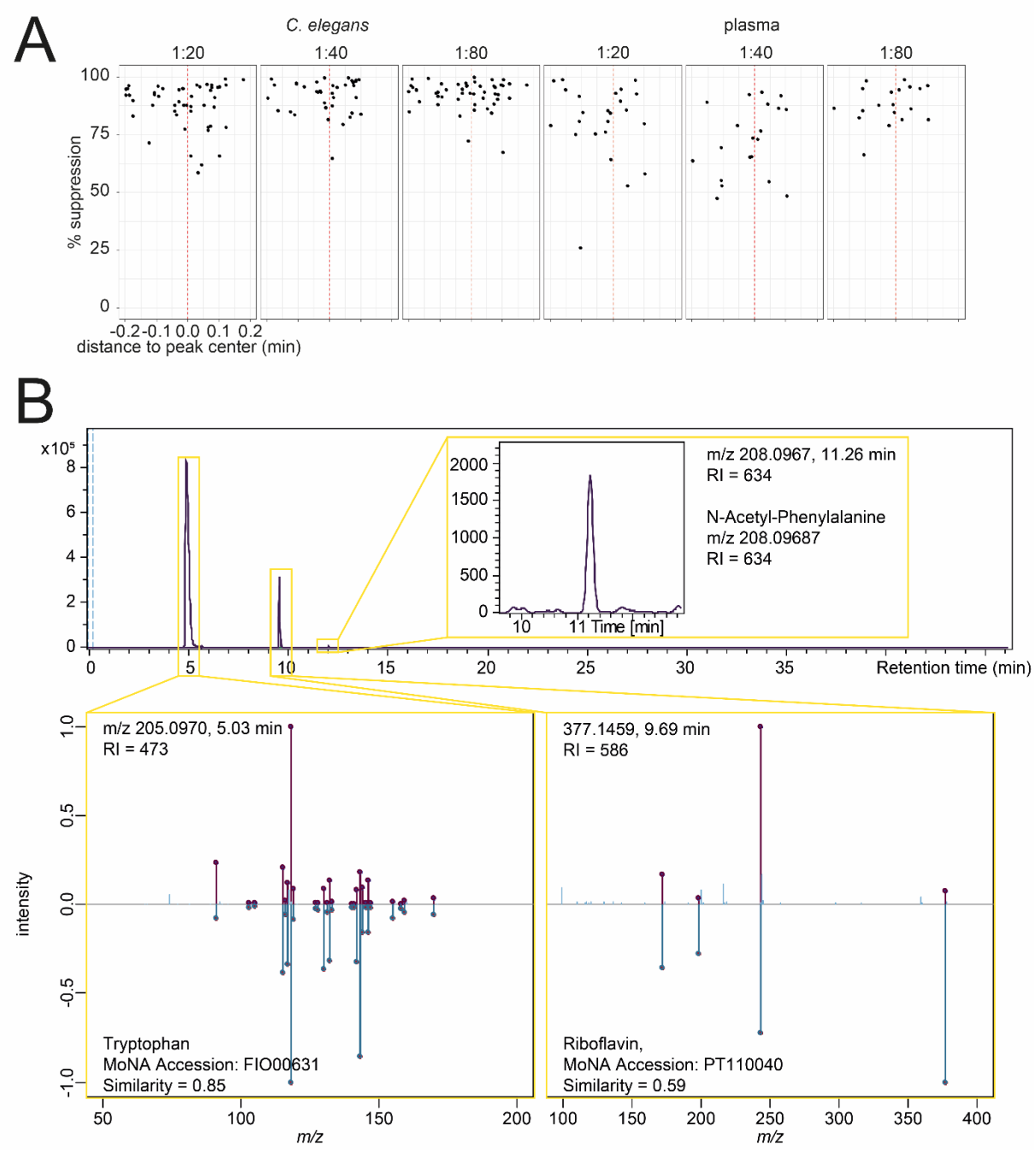


Figure 13: (A) Ion suppression effects of NAPS spiked into C. elegans or plasma metabolite extracts at different dilution factors were evaluated by comparison against unspiked matrix. The effect of suppression was checked in relation to the distance of a metabolite feature to the apex of the closest NAPS standard. The example shows the % suppression between -0.2 and 0.2 mins from the apex of the C15-NAPS. (B) Selection of annotation results from positive ionization mode of detected C. elegans metabolites. First matching was performed on m/z and RI. MS2 spectra have been compared with library spectra from MoNA. N-Acetyl-Phenylalanine was only matched based on m/z and RI values.

Using only *m/z* values with an error of 0.005 Da, 108 features were putatively annotated with one or several metabolites from the database in positive ionization mode. This number is reduced to 40 if, additionally, the RI is used with a maximum error of 10 RI units. Some of the annotated features were examined in more detail. First, a peak at 5.03 min with *m/z*

205.0970 was annotated as the $[M+H]^+$ adduct of tryptophan. The RI in the initial databases were 475 and 474, respectively, and the RI in the *C. elegans* sample was 472, which reflects an error of 2 RI units or 0.4%. When comparing the RT obtained on the same column with the same gradient on the Agilent 1200 HPLC system, a difference of 5% was observed, while for the RT used in the initial DB construction using a Waters Acquity UPLC, a difference of 46% was found. This peak has been selected for fragmentation using the data-dependent acquisition of MS/MS, and we, therefore, compared also obtained MS^2 spectrum. Spectra from the standard found in the MSMLS and the peak detected in *C. elegans* showed a perfect match. An additional spectral similarity search was performed on the MassBank of North America (MoNA). The closest hit was indeed tryptophan measured on a similar instrument (MoNA Accession FIO00631). Cosine similarity was 0.85. The second example was a feature with *m/z* 377.1459 detected at 9.69 min annotated as $[M+H]^+$ of riboflavin. Likewise, checking MS/MS data that was available matched a fragmentation spectrum of riboflavin (MoNA Accession PT110040). Lastly, *m/z* 208.0967 at 11.26 min was annotated as *N*-acetyl phenylalanine. This peak was low in intensity, and no MS/MS collection was triggered. Beside the *m/z* value, RI was used for annotation, which increased the confidence of the identification.

All examples so far had only one putative annotation in the used database. However, RI shall be used as orthogonal information for metabolite identification, which allows the filtering of false positive annotations from *m/z* values alone and/or MS^2 spectra. Therefore, the annotated data was checked for detected features that initially showed multiple annotations based on m/z alone. A signal with *m/z* 139.0389 detected at 6.51 min was annotated as $[M+H]^+$ of either salicylic acid (2-hydroxybenzoic acid), 3-hydroxybenzoic acid or 4-hydroxybenzoic acid. All three are isomers with the same sum formula, $C_7H_6O_3$, and very similar fragmentation patterns, only differing in the abundance of different fragments. Therefore, differences in retention behavior are required to identify the correct isomer. In addition, the RI results were reduced to a single hit, 4-hydroxybenzoic acid, known to be present in *C. elegans*, where it is one of the building blocks of complex ascaroside signaling

molecules [136]. Identifications of 4-hydroxybenzoic acid and riboflavin could also be confirmed in the negative ionization mode using matching *m/z* and RI values. All results are summarized in Table 16.

Table 16: Results of metabolite annotation of selected features detected in C. elegans using the constructed RI database. Matching is based m/z and either RT or RI. Matching has been performed on the basis of m/z values

Ionization Mode	Feature	Annotation	Type	UPLC (Std DB)	HPLC
positive	*m/z* 205.0970 5.03 min RI 473	Tryptophan $[M+H]^+$	RT	2.71 (46%)	5.29 (5%)
			RI	475 (0.4%)	477 (0.8%)
	m/z 377.1459 9.69 min RI 586	Riboflavin $[M+H]^+$	RT	4.66 (52%)	12.31 (27%)
			RI	583 (0.5%)	595 (1.5%)
	m/z 208.0967 11.26 min RI 634	*N*-acetylphenylalanine $[M+H]^+$	RT	5.38 (52%)	13.8 (23%)
			RI	634 (0%)	637 (0.55)
	m/z 139.0389 6.51 min RI 513	4-hydroxy benzoic acid $[M+H]^+$	RT	3.4 (47%)	6.87 (5.5%)
			RI	512 (0.2%)	512 (0.2%)
		3-hydroxy benzoic acid $[M+H]^+$	RT	4.30 (33.9%)	9.8 (50.5%)
			RI	562 (9.5%)	557 (8.6%)
		Salicylic acid $[M+H]^+$	RT	4.30 (33.9%)	9.8 (50.5%)
			RI	562 (33.9%)	557 (50.5%)
negative	*m/z* 137.0250 6.53 min RI 513	4-Hydroxybenzoic acid [M-H]-	RT	3.4 (47%)	6.83 (4.5%)
			RI	512 (0.2%)	512 (0.2%)
	m/z 375.1306 9.69 min RI 586	Riboflavin [M-H]-	RT	4.65 (52%)	n.d.
			RI	583 (0.5%)	n.d.

3.2.3.2. Cross-column pilot study

RIs in GC-MS can be used across the same column from different vendors, e.g. 5%-phenyl-95%-dimethylpolysiloxane, because their stationary phase preparation is very similar and well-defined. [359]. However, these columns typically show very narrow specifications and are almost identical. In contrast to this, different LC columns show different selectivity, even within one type of column (e.g. USP L1 or C18 columns). Though the general separation mechanism is not changed, different secondary interactions can lead to differences in selectivity and local changes in elution order and retention times. In order to check how far RI can normalize for different columns, in the first step different C18 columns from Waters were compared (compare Table 14). All columns were of the same dimensions with 100 mm length and 2.1 mm inner diameter. Three different gradient lengths were used with each column. Ninety-nine metabolite standards were injected, and indexing was performed, as mentioned in the previous study. Separation was achieved on a Waters Acquity UPLC system coupled to a Bruker maXis (same system as LC System 1 above). Although all columns are nominally all USP L1 phases, they exhibit different selectivity, which is influenced by the solid support, carbon load, end-capping, etc.

In order to get an overview of how close the different columns are regarding their selectivity, the hydrophobic subtraction model was used [372]. This model described the physicochemical nature of a particular stationary phase by five parameters, which are specifically H parameter as a measure of the phase hydrophobicity, S^* as a measure of the resistance of the stationary phase to penetration by a solute molecule, A as a measure of the hydrogen-bond acidity of the phase, B as a measure of the hydrogen-bond basicity of the phase, and C as a measure of the interaction of the phase with ionized solute molecules. Together with the characteristics for a given solute (η as a parameter of the solute hydrophobicity, σ as a measure of the bulkiness of the solute molecule, β as a measure of hydrogen-bond basicity of the solute, α as a measure of the hydrogen-bond acidity of the solute, κ as a measure of the ionization state of the solute), they are related to the retention of the solute (k_x) and retention of a given reference compound (e.g. ethylbenzene) (k_{EB}).

$$log\left(\frac{k_x}{k_{EB}}\right) = \eta H - \sigma S^* + \beta A + \alpha B + \kappa C$$

The different columns can be compared using the similarity factor F_s calculated by the following equation. w represents weighting factors and standard values are w_H = 12.5, w_{s^*} = 100, w_A =30, w_B = 143, $w_{C2.8}$ = 83. An F_s value of close to 0 means similar selectivity, while bigger values indicate different selectivity.

$$F_s = \sqrt{\left(w_H(H_1 - H_2)\right)^2 + \left(w_{S^*}(S_1^* - S_2^*)\right)^2 + \left(w_A(A_1 - A_2)\right)^2 + \left(w_B(B_1 - B_2)\right)^2 + \left(w_{C_{2.8}}\left(C_{2.8_1} - C_{2.8_2}\right)\right)^2}$$

For a fair comparison between the column, only substances detected in all conditions were used. The mean was used between positive and negative modes if the individual RI values did not differ by more than 5 RI units, and only substances with a RI > 300 were considered. In total, only 16 compounds were commonly detected between all the measurements. Although this number is relatively low, it allowed a first comparison.

First, two batches of BEH C18 columns were compared. As already expected, the RT as well as the RI, are highly reproducible and can be used to normalize between different column batches. Second, using the 16 compounds, it could be observed that between column pairs with high F_S value, the differences in RIs also show a tendency toward higher values. One particular example is sphingosine on the BEH C18 and BEH CSH C18 columns. The CSH columns have an additional positive charge on the surface. While the NAPS is virtually unaffected by this, sphingosine retention decreases by 21.29% in the used acidic conditions sphingosine is positively charged, leading to lower retention. Another example is Tryptophan as an early eluting substance showing a decrease of retention by 8.51%, likewise positively charged. One point to potentially correct for differences between columns is an additional two-point correction according to the following equation:

$$C_x = S_1 + (M_x - M_1)\frac{(S_2 - S_1)}{(M_2 - M_1)}$$

With Cx as the corrected RI of a given analyte x, M_1 and M_2 the measured RI values of 2 internal standards, S_1 and S_2 the standard RI values of 2 internal standards and M_x the

uncorrected RI value of analyte x. Two-point correction with phenylpyruvic acid as early eluting and linoleyl ethanolamide as late eluting reference was carried out. However, this 2-point correction was not able to improve cross-column comparison. Investigating the data in more detail, a compound class-specific effect is visible. While both sphingosine and tryptophan show a decrease comparing BEH C18 to BEH CSH C18, myristic acid, ferulic acid, chorismic acid, 2-hydroxyphenyl acetic, and arachidonic acid show an increase. Therefore, changes in RIs are substance-class-specific. If normalization can be carried out for specific substance classes, this might increase accuracy. Two-point correction with 3-Hydroxyphenylacetic acid and Myristic acid partially improved accuracy, e.g. deviation 3-Methylhippuric was reduced from 3.34% to 0.38% or deviation of Arachidonic acid was reduced from 6.70% to 0.68%. However, the deviation of sphingosine was increased from -21.29% to -24.67%, and the deviation of tryptophan was increased from -8.51% to -9.05%.

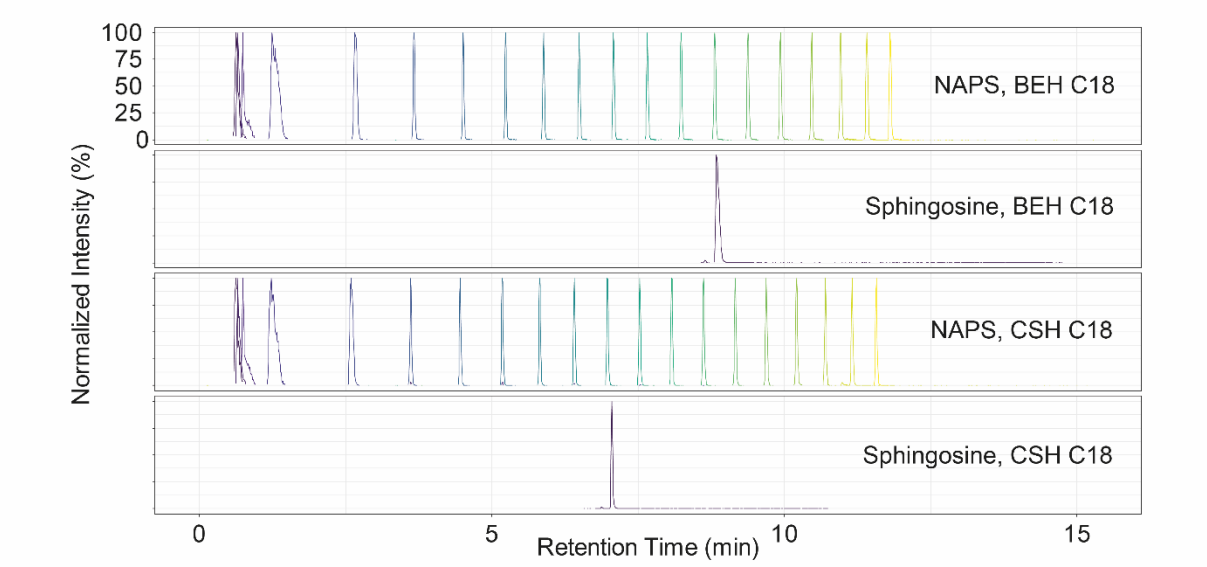


Figure 14: Comparison of retention of NAPS and Sphingosine on different columns. The upper two panels correspond to a Waters BEH C18 column, while the lower two correspond to a Waters CSH 18 column. While the retention of NAPS is virtually unaffected, sphingosine shows are large shift in retention and therefore also in the RI.

These results show that a cross-column comparison is only partially possible and only substance class specific. Since in non-targeted metabolomics, the identity and substance class is not known upfront, normalization of RT to RI is only useful for comparison within a single column chemistry. However, the database used is very small, and further

investigations are required to draw final conclusions. Likewise, the initial comparison between columns from different vendors showed only partial agreement, similar to the comparison described here.

3.2.3.3. Interlaboratory pilot study

After showing that RTI is useful for the normalization of retention information within one lab and a single separation chemistry, different laboratories shall be compared. This represents the typical use case one would expect for using an RI library to annotate metabolites in non-targeted metabolomics. Five different LC-MS systems in five different laboratories have been used for the evaluation of RIs to normalize differences between laboratories. The different LC systems are based on different pumping systems (LPG vs. HPG) and have different dead volumes.

The number of detected compounds in each laboratory ranged from 45 to 70. The number of pairwise overlapping compounds ranged from 33 to 59. To investigate how well the RIs matched between different laboratories, they were plotted against each other (Figure 15). For most laboratory pairs, a good fit has been observed. Errors were in the range between -20 and 20% relative deviation, with most of the error between -5 and 5%, which agrees to the first experiments from the intralaboratory pilot study. Larger errors might be attributed to wrong peak picking and are not further discussed. The only exceptions were found for laboratory 5, for which generally lower values were found. After discussion with the laboratory, it has been confirmed that they used a large injection loop, which caused a much larger dead volume and gradient delay volume compared to all other systems. Therefore, although the nominally the same gradient was programmed, the exact gradient conditions were very different. In order to make the data of laboratory five comparable, a 2-point correction as described above was used, using 4-Hydroxybenzoic acid and 25-Hydroxyvitamin D2 as the reference standard. After correction, values for laboratory 5 were comparable to others, with mostly 0-5% deviation, while for all other laboratories, no systematic shift was observed after correction. The 2-point correction didn’t affect the

performance in the case of the other laboratories and even improved the performance for certain metabolites.

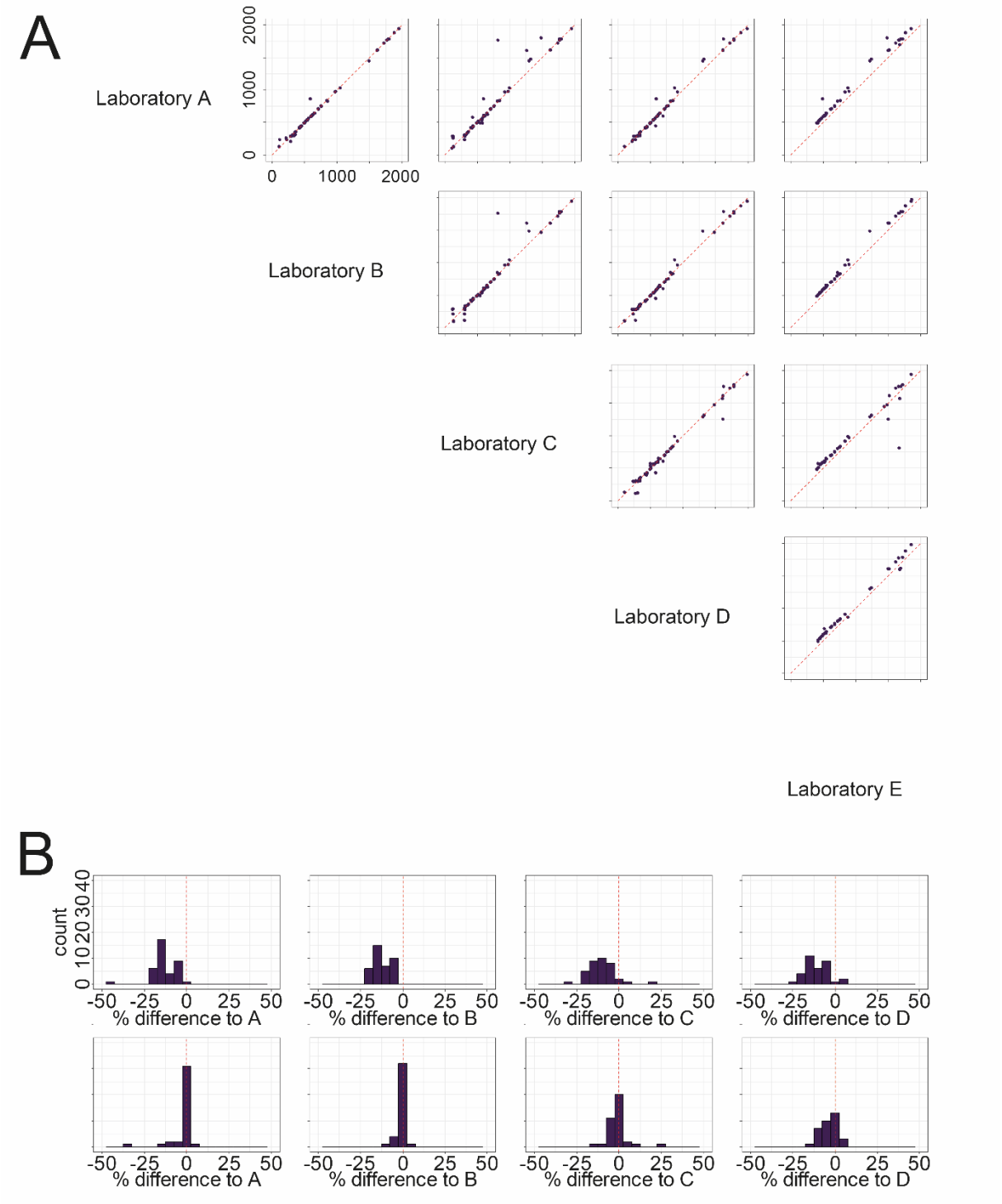


Figure 15: (A) Scatter plot comparing the obtained RI of metabolites between the different laboratories. Dotted halfline indicates the diagonal, which referes to a perfect match. Laboratory E shows a systematic deviation compared to all other laboratories (B) Histogram of relative differences of laboratory E to all other laboratories without 2-point correction (upper row) or with 2-point correction (lower row). 2-point correct was able to remove the systematic deviatoin present in the data.

3.2.3.4. Interlaboratory study

A larger ring trial was conducted based on the initial observation from the interlaboratory pilot study above that RTI can be used to normalize RT information between laboratories. The metabolites contained in the MSMLS used in the first study were used, but this time in a different column. RTs of compounds contained in this library on a Phenomenex Kinetex C18 column have been previously measured by Pezzatti *et al.* [373]

While in laboratory A pooled row mixes were measured, all other participating laboratories measured pooled mixes for the entire plates. After the injection of 5 blanks, one NAPS injection was followed by the seven mixes. The NAPS and the plate mixes were injected five times, followed by a final injection of the NAPS. RTIs were calculated by using the bracketing NAPS injections for the respective sample. Here the average of the RT of the respective NAPS standards from the injection before and after the standard mix was used for the calculation of the RI.

Laboratory A detected 419 out of 647, while laboratory B detected 320, laboratory C 371, and laboratory D 337 in positive and negative modes combined. If substances were detected in positive and negative ionization mode, the average of the calculated RIs was used if the individual values did not differ more than 10 RI units. After filtering for RI values > 300 and removing obvious outliers, 110, 69, 95, and 72 metabolites remained in laboratories A, B, C, and D, respectively. Similar to the pilot study, most errors were between -5 and +5%, indicating with this second independent study that comparison of RIs between different laboratories is possible.

In the next step, the possibility for annotation using RI cross laboratories was checked. Laboratory A measured a *C. elegans* metabolite extract, and annotation was performed based on the databases obtained from all other laboratories. Annotation was performed using the `MetaboAnnotation` package and an absolute m/z error of 0.005 Da and an RI error of 10 units. Data from the standards measured in parallel to the *C. elegans* extracts were used as ground truth and confirmed by MS^2 data. No secondary correction was

performed. Several metabolites could be identified using the databases from all laboratories, e.g. biotin. A sizeable systematic trend was found for laboratory B, which used a different LC system, a low-pressure quaternary pump. However, most of the systematic trends were removed by converting RT to RI. For most metabolites, the databases from the different laboratories yielded the same annotation, indicating that RIs can be used for cross-laboratory identification.

3.2.4. Conclusion

RT information represents valuable, orthogonal information to MS^1 and MS^2 data for metabolite identification. However, since no standard chromatographic setup exists the use of RT is restricted to single (mostly laboratory-specific) methods. Even between different laboratories using nominally the same chromatographic setup different RTs can be obtained due to factors like dead volume, pumping system, etc. RTI is used in GC to account, for example, for differences in column length. Using a mixture of NAPS homologs this approach was transferred to RP-based LC-MS and allowed to convert RT to RI and therefore, to normalize for different experimental setups. Based on the measurement of chemical reference standards, the approach was evaluated, and normalization could be achieved for differences in flow rate and temperature. Furthermore, separations were carried on different formats (UPLC vs. HPLC). This comparison showed very low deviations in the RIs, making it possible in the future to compare retention data from different systems using the same separation chemistry (same column and eluents). NAPS can be co-injected with the sample for the most accurate RTI, but ion suppression might be an issue. Obtained results indicate that at a sufficient dilution, NAPS are still detectable within biological matrices, but ion suppression effects are reduced. However, the dilution factor must be tested and optimized for each individual matrix and LC-MS setup. RIs can be used to improve the annotation and identification of metabolites by adding an additional orthogonal parameter to *m/z* and fragmentation pattern. Results have shown that isomeric species can be filtered RIs obtained on a different LC-MS setup.

The similarity of RTI values between different C18 columns was evaluated. All columns were used with ACN gradients and 0.1% formic acid as an additive. Nominally, all columns were classified as USP L1 columns (C18 column). However, the results were vastly different and some larger shifts in RT occurred RTI couldn't normalize that. A secondary correction could achieve improvements within single metabolite classes, but no global improvement could be observed. Individual correction is possible due to shared chemical properties within a single class, but the structural diversity of metabolites makes it not possible to find a good common normalization. Based on the results, it can be concluded that RIs cannot be used across different columns. Still, the number of tested columns is comparably small to the number of available C18 columns. Columns that are very similar on the basis of the hydrophobic subtraction model might lead to more similar results.

Still, comparison of RIs within one separation chemistry is possible, which is not only possible within a single laboratory but also allows the normalization of RT information between different laboratories. An initial ring trial conducted between 5 laboratories showed good agreement of RIs between the laboratories using similar separation systems. A second larger ring trial confirmed these results using a different column and a larger set of metabolites. The usability of RI for cross-laboratory annotation was checked by annotating metabolites from *C. elegans* metabolite extracts, indicating comparable results between the databases from the different laboratories.

Based on the findings, it can be concluded that RTI in reversed-phase LC-MS will help to share RT information between laboratories and help to standardize the reporting. The current databases generated are of limited size, and in the future, they have to be populated with additional data on further metabolites and separation conditions. It might be risky to use *m/z* and RI values alone for metabolite identification. Instead, matching of RIs shall be used to re-rank results from tandem MS search or combined with accurate mass and tandem MS matching in an integrated (consensus) scoring function. So far, RTI has been only evaluated for C18 columns. If this or a similar approach will work for HILIC-based separations remains

to be elusive. Even if C18 columns already show some deviations, the variety of HILIC columns is much larger with different stationary phases and selectivities. All initial results show that conversion of RT to RI can help to normalize data to allow cross-laboratory comparison without the need to calculate mapping functions such as PredRet is using [323]. For example, MassBank already supports the NAPS-based RI in their database.

3.3. Ion mobility information

Ion mobility separation (IMS) is increasingly integrated into metabolomics and lipidomics workflows. While RT represents a system property dependent on the metabolites of interest, stationary and mobile phase, and experimental parameter employed, CCS represents a molecular property and only shows minor variation across different ion mobility systems. Different types of ion mobility instruments have been used in metabolomic and lipidomics. One of the first applications was described by Paglia *et al.* [374]. CCS values were employed on top of RT for the annotation of metabolomics from epithelial and mesenchymal cells. CCS were determined for 125 metabolites using TWIMS, and an RSD of < 2% for the CCS values was observed. Using the stepped-field method, Zheng et al. used DTIMS to measure CCS values for over 500 small molecules [375]. Different examples showed the possibilities for separation of molecules, e.g. α- and β-muricholic acid had very similar drift times while the isobaric molecules tyrosine and glufosinate were baseline separated in positive and negative ionization mode. Lastly, confidence in annotation on the MS^1 level could be improved by combining accurate mass search with CCS matching.

However, to be useful for metabolite identification, CCS values must be comparable between different laboratories and instruments. Different cross-laboratory comparisons have been performed, e.g. Stow *et al.* performed a comparison of substances from different classes using DTIMS [376]. Results from four laboratories were compared, and errors were within 1.5% absolute percent error for all classes. Nye *et al.* performed a comparison of TWIMS instruments located at two different sites [377]. They found good agreement of $^{TW}CCS_{N2}$ values between the instruments in direct infusion and LC separation hyphenated to IM-MS. Errors within 1% for DI-IM-IMS and +/- 2% for UHPLC-IM-MS were found. However, both publications compared CCS values within a single type of instrumentation. In contrast to this, Hinnenkamp *et al.* compared $^{DT}CCS_{N2}$ and $^{TW}CCS_{N2}$ values for 124 substances [378]. Mean deviations of 1.0% for $[M+H]^+$ and 1.1% for $[M+Na]^+$ ions were found, but also deviations up to 6.2% were observed.

While the deposition of MS^2 data in repositories such as MassBank or GNPS is becoming more common, only a few databases for CCS values are available. One of the most prominent examples is the CCS compendium storing information independent of the used instrumentation [379]. Likewise, the websites for the CCS prediction tools CCSBase and AllCCS, have associated databases used as training datasets, which can also be searched [350, 380]. Likewise, similar resources have been built for lipids [381, 382]. Ion mobility and CCS are particularly interesting for the annotation of lipids. Based on their structure, they are large enough to show a difference in their confirmation. Trendlines along the plot of CCS vs. *m/z* can be used to filter false positive annotations. Additionally, potential isomers can be separated if the IMS resolution is large enough.

In this work, the usability of DTIMS for the analysis of small molecules as well as lipids is examined. Since ion mobility represents a second separation dimension leading to less crowded MS data, the possibility of combining IMS and DIA for lipid analysis was explored as well.

3.3.1. Material and Methods

3.3.1.1. Chemicals

HPLC-grade methyl-tert-butyl ether (MTBE) and LC-MS-grade methanol (MeOH), iso-propanol (iPrOH), acetonitrile (ACN), ammonium formate, and formic acid were obtained from Sigma-Aldrich. Water was purified using a Merck Millipore Integral water purification system with a resistance of 18 MΩ and TOC < 5 ppb. Reference standards were purchased from different chemical vendors (e.g. Sigma-Aldrich, Merck, Cayman, etc.) and dissolved in an appropriate solvent. Aliquots of stock solutions were stored at -20°C until further analysis. Lipid reference standards were purchased from Avanti Polar lipids and are summarized in SI Table 3. Maradolipid standards have been synthesized by the Knölker group [383]. A mix standard consisting of 6-*O*-myristoyl-6'-*O*-myristoyltrehalose (Mar 14:0/14:0), 6-*O*-(13-methylmyristoyl)-6'-O-(13-methylmyristoyl)trehalose (Mar 15:0/15:0), 6-O-myristoyl-6'-O-oleoyltrehalose (Mar 14:0/18:1), 6-O-palmitoyl-6'-O-palmitoyltrehalose (Mar 16:0/16:0), 6-O-

(13-methylmyristoyl)-6'-O-(15-methylpalmitoyl)trehalose (Mar 15:0/17:0), 6-O-(13-methylmyristoyl)-6'-O-oleoyltrehalose (Mar 15:0/18:1), 6-O-palmitoyl-6'-O-oleoyltrehalose (Mar 16:0/18:1), 6-O-(15-methylpalmitoyl)-6'-O-oleoylterhalose (Mar 17:0/18:1), 6-O-oleoyl-6'-O-oleyoltrehalose (Mar 18:1/18:1) and 6-O-oleoyl-6'-O-(2-octyl-cyclopropaneoctanoyl)trehalose (Mar 18:1/19:1) was dissolved in MeOH.

3.3.1.2. Stepped field method lipids

CCS of lipid standards and maradolipid standards were collected using the stepped field method by Stow *et al.* using an Agilent 6560 DT-IM-Q-ToFMS equipped with a Dual Agilent Jet Stream ESI source (Agilent Technologies, Waldbronn, Germany) [376]. Ionization source parameters were as follows: positive ionization mode: Vcap 4000 V, Nozzle Voltage 2000 V, Fragmentor 400 V, Gas Temperature 250°C, Gas Flow 12 L/min, Nebulizer 40 psig, Sheath Gas Temperature 320 °C, Sheath Gas Flow 11 L/min; negative ionization mode: Vcap 5500 V, Nozzle Voltage 2000 V, Fragmentor 400 V, Gas Temperature 250°C, Gas Flow 12 L/min, Nebulizer 40 psig, Sheath Gas Temperature 320 °C, Sheath Gas Flow 11 L/min. The instrument was operated with N_2 as drift gas at a pressure of 3.95 Torr. The arrival time of ions in the IMS-Q-ToF-MS represents the sum of the time spent inside and outside the drift region. To accurately determine the correct drift time ($t_d = t_D - t_0$), the electrical field is varied. This leads to changes in t_D, since t_0 is constant, the varying part represents the actual drift time. By plotting t_D over 1/V t_0 can be determined as an intercept. Electrical fields were varied by keeping the rear funnel at constant potential and varying the drift tube entrance voltage.

The maradolipid standard mix was diluted in a 50/50 mixture of eluent A and eluent B (see below) and infused using a syringe pump with a flow rate of 500 µL / min. All other lipid standards were diluted in iPrOH / $CHCl_3$ / MeOH (4/2/1, v/v/v) with 7.5 mM ammonium formate and infused using a syringe pump. In negative ionization mode, a flow rate of 1000 µl/h and in positive ionization mode, 500 µl/h were used. Each lipid standard was infused three times, and the average was calculated. Data was analyzed using the Agilent MassHunter Workstation IM-MS Browser 10.0. IM data has been referenced using either

using the $[M+H]^+$ adduct of Hexakis(1H,1H,3H-perfluoropropoxy)phosphazene (*m/z* 922.0098, $^{DT}CCS_{N2}$ 243.64 Å²) or the $[M+TFA-H]^-$ adduct of Hexakis(1H,1H,3H-perfluoropropoxy)phosphazene (*m/z* 1033.9870, $^{DT}CCS_{N2}$ 255.34 Å²) from the reference mass solution in positive and negative ionization mode respectively.

3.3.1.3. C. elegans cultivation and extraction

daf-2(e1370) mutants were obtained from the Caenorhabditis Genetics Center (CGC) and grown under standard conditions according to Brenner *et al.* on Nematode Growth Medium (NGM) [4]. To obtain dauer larvae, synchronized L1 larvae were obtained by bleaching and seeded onto NGM plates and grown at 25°C. Once sufficient amounts of dauer larvae were obtained, worms were washed off the plates using an M9 buffer and washed three times. Lipids were extracted, according to Bligh and Dyer [384]. The chloroform phase was evaporated to dryness and redissolved in H_2O/ACN/iPrOH (5/35/60, v/v/v) prior to analysis.

3.3.1.4. LC-IMS-MS, single field method lipids

Lipid analysis was performed using two different methods. All lipids standards were analyzed with the method published by Witting *et al.* [68]. Briefly, lipids were separated on Waters CORTECS UPLC C18 column (150 mm x 2.1 mm ID, 1.6 µm particle size) using a linear gradient from eluent A (40% H_2O / 60% ACN + 10 mM ammonium formate / 0.1 % formic acid) to eluent B (10% ACN / 90% iPrOH + 10 mM ammonium formate / 0.1% formic acid). The following gradient was used: 68/32 at 0 min, 68/32 at 1.5 min, 3/97 at 21 min, 3/97 at 25 min, 68/32 at 25.1 min with a flow rate of 0.250 mL/min and a temperature of 40°C. The column was re-equilibrated for 2.5 minutes.

Additionally, all lipids except the maradolipids were analyzed according to the LC method adopted from Knittelfelder *et al.* [385]. Briefly, solvents were similar, but a different column was employed. Lipid separation was performed using an Agilent ZORBAX RRHT Extend C18 column (2.1 x 50 mm, 1.8 µm, Agilent Technologies, Waldbronn Germany). Eluent A consisted of 100% water + 10 mM ammonium acetate / 0.1% formic acid / 8 µM phosphoric acid and eluent B of 100% iPrOH + 10 mM ammonium acetate / 0.1% formic acid / 8 µM

phosphoric acid. The following gradient conditions were used: 60/40 at 0 min and 0.2 mL/min, 60/40 at 1.00 min and 0.2 mL/min, 60/40 at 1.05 min and 0.4 mL/min, 60/40 at 1.5 min, 40/60 at 3.0 min, 12/88 at 12.0 min, 0/100 at 14.50 min, 0/100 at 15.25 min, 60/40 at 15.30 min, 60/40 at 15.62 min, 60/40 at 15.70 min and 0.2 mL/min, 60/40 at 16.0 min at 0.2 mL/min.

Detection was performed in positive and negative ionization modes using an Agilent 6560 IMS-QToF-MS equipped with a Dual-Jetstream ESI source (Agilent Technologies, Waldbronn, Germany). A reference mass solution was infused using an additional isocratic pump. Lipid separation was performed on an Agilent 1290 Infinity II UHPLC. At the beginning of each measurement batch, the ESI-L tune Mix was injected for 1 minute for CCS calibration.

3.3.1.5. LC-IMS-MS, single field method metabolites

Metabolites were analyzed on a Phenomenex Kinetex C18 column (150 mm x 2.1 mm ID, 1.7 µm particle size) using a linear gradient from eluent A (100% H2O + 0.1% formic acid) to eluent B (100% ACN + 0.1% formic acid). The following gradient was used: 98/2 at 0.0 min, 0/100 at 14 min, 0/100 at 17 min, 98/2 at 17.1 min with a flow rate of 0.300 mL/min and a temperature of 30°C. The column was re-equilibrated for 7.9 minutes. Detection was performed in positive and negative ionization modes using an Agilent 6560 IMS-QToF-MS equipped with a Dual-Jetstream ESI source (Agilent Technologies, Waldbronn, Germany). A reference mass solution was infused using an additional isocratic pump. Lipid separation was performed on an Agilent 1290 Infinity II UHPLC. A the beginning of each measurement batch 1 minute of ESI-L tune Mix was injected for CCS calibration.

3.3.2. Results

3.3.2.1. Theoretical considerations

CCS values, in contrast to MS^1, MS^2, MS^n, and RT, are not covered in current metabolite identification schemes, e.g. MSI levels or levels according to Schymanski *et al.* [281, 297, 386]. First, ion mobility instruments are not as widely used as other instrumentation, despite

the fact they were introduced some time ago. While RT, for example, represents a system's property resulting from the combination of analyte, stationary and mobile phase as well as the instrumentation, CCS is nearly independent of the instrumentation. However, slight variations between different instrument types can be observed. Generally, DTIMS and TIMS are in good agreement, while TWIMS shows some systematic offset. This offset is typically in the range of a few percent. Causon and Hann derived from some theoretical considerations that a minimal difference in CCS values of 1.5 to 1.8%for a resolution R_s = 0.6, while a baseline separation with a resolution R_s = 1.5 requires 3.7 to 4.4% [387].

In order to evaluate the usefulness of CCS values, metabolites in the CCS compendium were compared based on their *m/z*, CCS, and logP as a proxy for chromatographic retention time [379]. Based on structural information, the XlogP was calculated using the R package `rcdk` [388]. Since in metabolomics and lipidomics, metabolites are measured using a combination of IMS and MS, only *m/z* values that fall in a window of 0.01 Da were compared against each other. m/z values with corresponding charges between -3 and +3 were considered. In total, 2344 and 1231 metabolites fulfilled this criterion in positive and negative ionization modes, respectively.

In positive and negative ionization modes all possible pairs with a mass difference < 0.5 Da were examined. Since CCS shall be used as an orthogonal identifier for metabolite identification, it should help to distinguish between isomeric and isobaric structures. In positive ionization mode, 4204 and in negative ionization mode, 2284 pairs were examined. Differences in CCS were expressed as relative differences. In negative mode, differences up to almost 80% were observed. However, this difference was observed for an *m/z* difference over 0.2 Da. Positive ionization mode yielded differences over 100%, but only for higher mass differences. Therefore, the focus was on pairs with a mass difference smaller than 0.01 Da. Data were categorized into distinct groups for the *m/z* and CCS differences. In both cases, positive and negative ionization mode pairs fall into the category with an m/z difference < 0.005 Da and a relative CCS difference < 2.5 % (61.3% and 79.6 % of all pairs

in positive and negative mode, respectively). One particular example of great interest is the pair L-Leucine and L-Isoleucine. These molecules have the same chemical formula and the same mass. Additionally, their fragmentation spectra are almost identical. Therefore, orthogonal data such as RT and CCS are essential. However, in positive and in negative ionization modes, differences in CCS were below 2% in most cases. In contrast, they have a slight difference in their logP and some chromatographic methods can baseline resolve these isomers.

In contrast to MS^2 and RT, CCS can be predicted relatively simply. Different tools for this task have been described. Mobcal is one of the premier tools that has been used for the calculation of CCS values [389, 390]. This software has also been improved for higher throughput and more accurate predictions [391]. The new ISiCLE workflow refactors some of the Mobcal functionalities and enables improved predictions [348]. However, *ab initio* prediction as performed by this software still takes a long time, often not compatible with the time scale of metabolomics and lipidomics experiments, where several hundreds or thousands of CCS values for potential lipids need to be predicted. Several tools for the prediction of CCS from molecular descriptors have been described. One of the first tools was MetCCS [392]. This tool allows the prediction of CCS values from 14 molecular descriptors, including the *m/z* value, which is one of the most important descriptors. A median relative error of 3% was achieved. A web service for prediction has been established as well for custom predictions [351]. A version specific to the prediction of lipid CCS values has been developed by the same group with an increased median relative error of ~1% [393]. Lastly, CCSbase has been described using molecular quantum numbers as structural characteristics together with machine learning for CCS prediction [350].

3.3.2.2. Ion mobility and CCS values for metabolite identification

Though the use of CCS might be limited, it was to test if they can be used for metabolite identification. Two different experiments were conducted, one within the domain of metabolites and one for lipids. First metabolites from the previously mentioned RI

interlaboratory study (see chapter 3.2) have been measured using a DTIMS instrument, and besides RT also, the CCS for the $[M+H]^+$ and $[M-H]^-$. In order to use the CCS value for metabolite identification, a reference database for CCS values is required. The employed Agilent 6560 IMS-Q-ToF-MS uses a DT and allows the determination of absolute CCS values experimentally without the need for calibration. As a second instrument, a TWIMS, has been used by the group of Serge Rudaz. One condition yielded CCS values for 211 metabolites and the second for 322. The overlap yielded 182 metabolites. While for the calculation of the RI, only metabolites with sufficient retention were evaluated, CCS values from all detected metabolites were used for comparison of the two instruments. Generally, a good agreement was found. However, a slight systematic deviation was found. $^{TW}CCS_{N2}$ had, on average, a 2-3% small value compared to $^{DT}CCS_{N2}$. Such behavior is well known since the results of CCS determination in TWIS are dependent on the used calibrant, which can lead to systematic deviations [394]. For all metabolites covered in the MSMLS library, CCS values were predicted using CCSbase and AllCCS [350, 380]. For comparison of both prediction tools against the DTIMS, a good agreement was found with most of the relative deviations between -5 and +5%, though a trend towards higher predicted CCS in AllCCS was observed. The same results were observed when comparing TWIMS-derived CCS values. Generally, CCS is an additional parameter addition additional confidence in identification but does not have the power to improve metabolite identification. The resolution and accuracy of the current instrumentation are too low, and therefore, CCS often can often remove only obvious wrong annotations but does not point toward the correct annotation. Therefore, the use of CCS in metabolite identification is not further discussed.

3.3.2.3. Ion mobility and CCS values for lipid identification

A lipid CCS reference data set for the employed LC-IMS-MS setup was collected from authentic lipid standards from Avanti Polar lipids. CCS values from lipids have been determined using different analytical strategies. First, the multifield or stepped-field method was used as a reference method, and lipids were directly infused using a syringe pump. In parallel, two different LC-IMS-MS methods have been used. The first method represents a

reference method supplied by Thomas Eichmann, and measurements have been performed in the frame of an interlaboratory ring trial on lipid $^{DT}CCS_{N2}$ values (unpublished). This method uses a short column and a faster gradient and can separate lipid species only partially, while the second method, according to Witting *et al.*, uses a longer column and separation time [68, 385]. Table 17 shows results from the measurement of PCs in positive mode as an example. Other values can be found in SI Table 4 and SI Table 5.

As the first step, the deviation between the stepped and single measurements was determined. Generally, $^{DT}CCS_{N2}$ values from the different methods were in good agreement. However, a higher variation was observed for $[M+H]^+$ and $[M+Na]^+$ adducts of PCs. While variations for all other lipid classes generally were in the range of an error of 0.8%, PCs showed higher deviations of 1% and above. $^{DT}CCS_{N2}$ calibration of the system was correct, and the $^{DT}CCS_{N2}$ value of *m/z* 922.0098 from the reference mass mix was determined correctly. A potential reason for the higher deviation might be overfilling of the trapping funnel before the ion mobility separation. Since all other used lipid classed values matched well between the multi- and single-field methods, it can be concluded that single-field measurements are usually sufficient for determining accurate $^{DT}CCS_{N2}$ values, as shown before [376]. Values between the two different single-field methods obtained from LC-IMS-MS showed a good agreement.

In the next step, CCS values between different instruments were compared. Values from the University Graz were provided by Thomas Eichmann and used for comparison (personal communication, unpublished). They have been determined using LC-IMS-Q-ToF of reference lipid extracts. Lipids were identified based on MS^2 data. Therefore, no detail on the sn1 and sn2 position as well as specific fatty acid composition or stereochemistry of double bonds, is known available for these measurements. Since one set of single field $^{DT}CCS_{N2}$ values were obtained with the same LC method as in Graz RTs can be used as additional matching criteria, although isomers are generally eluting close to each other. Additionally, also using

some tolerance for RT is often still multiple isomers matched. On average, an RT offset between 0.2 and 0.3 minutes was observed.

Since several included reference standards represent isobaric species the possibility for separation in the chromatographic and ion mobility domain was examined. Examples are PC 16:0/18:0, PC 18:0/16:0, and PC 17:0/17:0. For the three reference standards, the following $^{DT}CCS_{N2}$ values have been obtained respectively: 292.80 Å^2, 292.20 Å^2 and 292.40 Å^2 and 7.56 min, 7.54 min, and 7.54 min. Such values render the separation with the short method virtually impossible. Differentiation was only possible because lipids were injected independently from each other. Likewise, also the longer separation method yields identical retention times. Other examples are PC 16:0/18:1(9Z) and PC 18:1(9Z)/16:0 and PC 18:0/18:1(9Z) and PC 18:1(9Z)/18:0. While the short LC method again yields the same retention time, the longer method is able to separate the first pair with 0.2 minutes difference.

Table 17: Comparison of DTCCSN2 values obtained using the stepped field and single field methods with two different LC-IMS-MS methods. Values are obtained from the method also used by Thomas Eichmann and by the LC-MS method from Witting et al. [68].

Lipid	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield (Å^2)	$^{DT}CCS_{N2}$ Single Field (Å^2, Eichmann *et al.*)	$^{DT}CCS_{N2}$ Single Field 2 (Å^2, Witting *et al.*)
PC 14:0/14:0	$[M+H]^+$	678.5068	271.42	274.40 (1.10%)	273.33 (0.70%)
PC 15:0/15:0	$[M+H]^+$	706.5381	277.67	280.60 (1.06%)	278.93 (0.46%)
PC 16:0/16:0	$[M+H]^+$	734.5694	283.54	286.80 (1.15%)	285.43 (0.67%)
PC 16:0/18:0	$[M+H]^+$	762.6007	289.51	292.80 (1.14%)	291.97 (0.85%)
PC 16:0/18:1(9Z)	$[M+H]^+$	760.5851	286.74	289.60 (1.00%)	289.50 (0.96%)
PC 16:0/18:2(9Z,12Z)	$[M+H]^+$	758.5694	283.79	287.10 (1.17%)	285.87 (0.73%)

Lipid	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield (Å^2)	$^{DT}CCS_{N2}$ Single Field (Å^2, Eichmann *et al.*)	$^{DT}CCS_{N2}$ Single Field 2 (Å^2, Witting *et al.*)
PC 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+H]^+$	782.5694	287.37	290.10 (0.95%)	289.07 (0.59%)
PC 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+H]^+$	806.5694	289.93	292.90 (1.02%)	291.37 (0.50%)
PC 17:0/17:0	$[M+H]^+$	762.6007	289.57	292.20 (0.91%)	291.43 (0.64%)
PC 18:0/16:0	$[M+H]^+$	762.6007	289.64	292.40 (0.95%)	291.67 (0.70%)
PC 18:0/18:0	$[M+H]^+$	790.6320	295.41	298.80 (1.15%)	297.20 (0.61%)
PC 18:0/18:1(9Z)	$[M+H]^+$	788.6164	292.70	295.90 (1.09%)	295.47 (0.94%)
PC 18:0/18:2(9Z,12Z)	$[M+H]^+$	786.6007	290.16	293.30 (1.08%)	293.17 (1.04%)
PC 18:0/20:4(5Z,8Z,11Z,14Z)	$[M+H]^+$	810.6007	293.78	296.30 (0.86%)	296.80 (1.03%)
PC 18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+H]^+$	834.6007	296.68	299.70 (1.02%)	299.93 (1.10%)
PC 18:1(9Z)/16:0	$[M+H]^+$	760.5851	287.37	290.00 (0.92%)	299.17 (1.32%)
PC 18:1(9Z)/18:0	$[M+H]^+$	788.6164	293.32	296.20 (0.98%)	297.03 (1.26%)
PC 20:0/20:0	$[M+H]^+$	846.6946	307.80	310.80 (0.97%)	310.67 (0.93%)
PC 22:0/22:0	$[M+H]^+$	902.7572	319.01	---	321.47 (0.77%)

For several lipids present in the obtained dataset, reference values have been deposited in the CCS compendium [379]. For all species with detailed side chain composition available

good agreement between the measured values and the reference values was obtained in both ionization modes. LipidCCS was used to perform the prediction of CCS values for all lipids [393]. In positive mode very good agreement between the predictions for $[M+H]^+$ and $[M+Na]^+$ was found with an average relative error of ~0.3%. Error in negative mode was higher, with an average relative error of ~-1.45%. The higher error might be due to lower prediction performance because, generally, less training data is available in negative ionization mode.

3.3.2.4. CCS database creation for C. elegans lipids

Only a few lipids are available as standards and can be therefore used for the determination of CCS values. Since the prediction of CCS values using LipidCCS matched well with the measured CCS values for all predicted and supported lipids, CCS values were predicted [393]. Lipids classes for which a prediction was possible included: LPAs, PAs, MGs, DGs, TGs, PCs, LPCs, PEs, LPEs, PSs, LPA-Os, PA-Os, DG-Os, PE-Os, LPE-Os, PE-Ps, LPE-Ps, PC-O and LPC-Os. Prediction of lipids with one or more branched chain fatty acids was not supported by LipidCCS, yielding 272.7 Å² as a result. CCS values of lipids producing this predicted value were set to 0.

The predicted CCS values were used to investigate the possibilities for the use of CCS values in lipid identification, for which different scenarios of isomerism need to be discussed. The first case are sn-positional isomers, e.g. PC 16:0/18:1(9Z) and PC 18:1(9Z)/16:0. Second, double bond positional isomers exist like PC 16:0/18:1(9Z) and PC 16:0/18:1(11Z) and third case are cis/trans isomers, for example, PC 16:0/18:1(9Z) and PC 16:0/18:1(9E). Besides isomers, isobaric structures might exist, e.g. lipids from different classes having the same molecular formula (PC 16:0/18:1(9Z) and PE 17:0/20:1(11Z)) or PC 33:1 ($[M+FA-H]^-$) and PS 36:0 ($[M-H]^-$) in negative ion mode. Furthermore, isobaric masses of different adducts have to be deciphered. For example, the $[M+Na]^+$ ion of PC 34:1 is close to the $[M+H]^+$ ion mass of PC 36:4. The mass difference equals roughly the mass difference between two CH_2 groups and three double bonds ($\Delta m/z$ = 0.002).

In the next step all possible combinations of lipids from either only PCs (ionized as $[M+H]^+$ or $[M+Na]^+$), only PEs or PEs and PCs were examined in positive ionization mode while PC and PS were examined in negative ionization mode. In all further discussion, an isomeric overlap is defined as lipids that have exactly the same *m/z* value and isobaric overlap as lipids that have an *m/z* difference smaller than 0.02, which equals about the mass difference that can be separated with a resolution of R_s = 40.000 in mass range for lipid analysis. For the isomeric overlap within PCs and PEs individually, but also the isomeric overlap between PE and PC, the relative difference between the CCS values (ΔCCS) was always below 1.5%. In the case of the isobaric overlap, some pairs exceeded the 1.5%, but never the 4.5% required for baseline separation as defined by Causon and Hann [387]. Additionally, to the CCS values, the logP value was predicted and was used as a proxy for the difference in retention. While only small differences between the isomeric and isobaric overlap within the same class were observed, large differences were found between the lipid classes. In the case of the negative ionization mode isomeric overlap between PCs ionized as $[M+FA-H]^-$ and PS ionized as $[M-H]^-$ more pairs exceeded the 1.5% difference in CCS but also never exceeded the 4.5% threshold. In the case of logP large differences were observed. Differences in logP values arise from the difference in the headgroups and chain length. PC headgroups have a permanently charged headgroup which increases polarity compared to PE species of similar elemental composition. This is also partially in agreement with the findings of Blazenovic *et al.* [395]. Their work used different classification models based on *m/z*, RT, and CCS in different combinations to predict the lipid class or the lipid class and carbon number. The highest scores were achieved when combining *m/z*, RT, and CCS, indicating that RT is valuable information for lipid identification.

Since the observation so far is based, predicted values measured CCS values of lipids identified from *C. elegans* samples using LC-IMS-MS were compared. Identification of lipids was performed using obtained MS/MS spectra matched against a publicly available version of LipidBlast and inspected manually. All possible pairs of positive identifications were generated and filtered for a mass difference small 0.02 Da, similar to the predicted data. The

relative difference between the measured CCS values (ΔCCS) and the absolute difference in RT was calculated and compared. The pairs were grouped into isomeric and isobaric overlap and intra- or inter-lipid classes based on the identifications. Since the reported lipid features were already grouped according to adducts in the positive ionization mode, the table only contained $[M+H]^+$ adducts as main features. Therefore, only the isomeric overlap within or between lipid classes could be evaluated. In total 435 non-duplicate pairs were examined in positive ion mode. Out of these, 209 were pairs within the same lipid classes. Similar to the predicted values, the majority of the pairs had a ΔCCS small than 1.5%, and no pair showed values higher than 4.5%. However, retention time differences ranged from 0.1 to 2 minutes for most pairs. For the pairs between different lipid classes, the difference in CCS values spread between 0 and 2.3%, but RT differences of mostly larger than 1 minute were observed. In negative ionization mode, 100 pairs, of which 96 were within the same lipid classes, were examined. The results mirrored the positive mode results, and all ΔCCS values of the pairs were below 1%, while the RT difference ranged from 0.1 to 2.5 minutes. These results suggest that m/z and CCS values alone are not sufficient for the preliminary annotation of lipids and do not improve annotation problems for isomeric and isobaric overlaps. RT adds an important additional dimension to the correct putative annotation.

Another use of the CCS value instead of single values is to create trend lines along lipids of the same class. Groessl *et al.* used such trendlines to show that lipids that are very close in *m/z* (differing by 0.03 Da), e.g. PC-Os and PE can be grouped according to such lines [396]. Investigating such trendlines for different lipid classes examined no significant difference in slope could be found. However, lipids of the same class were generally closer to the line of their own lipid class compared to the overlapping class. Therefore, such trendlines might be useful for manual interpretation.

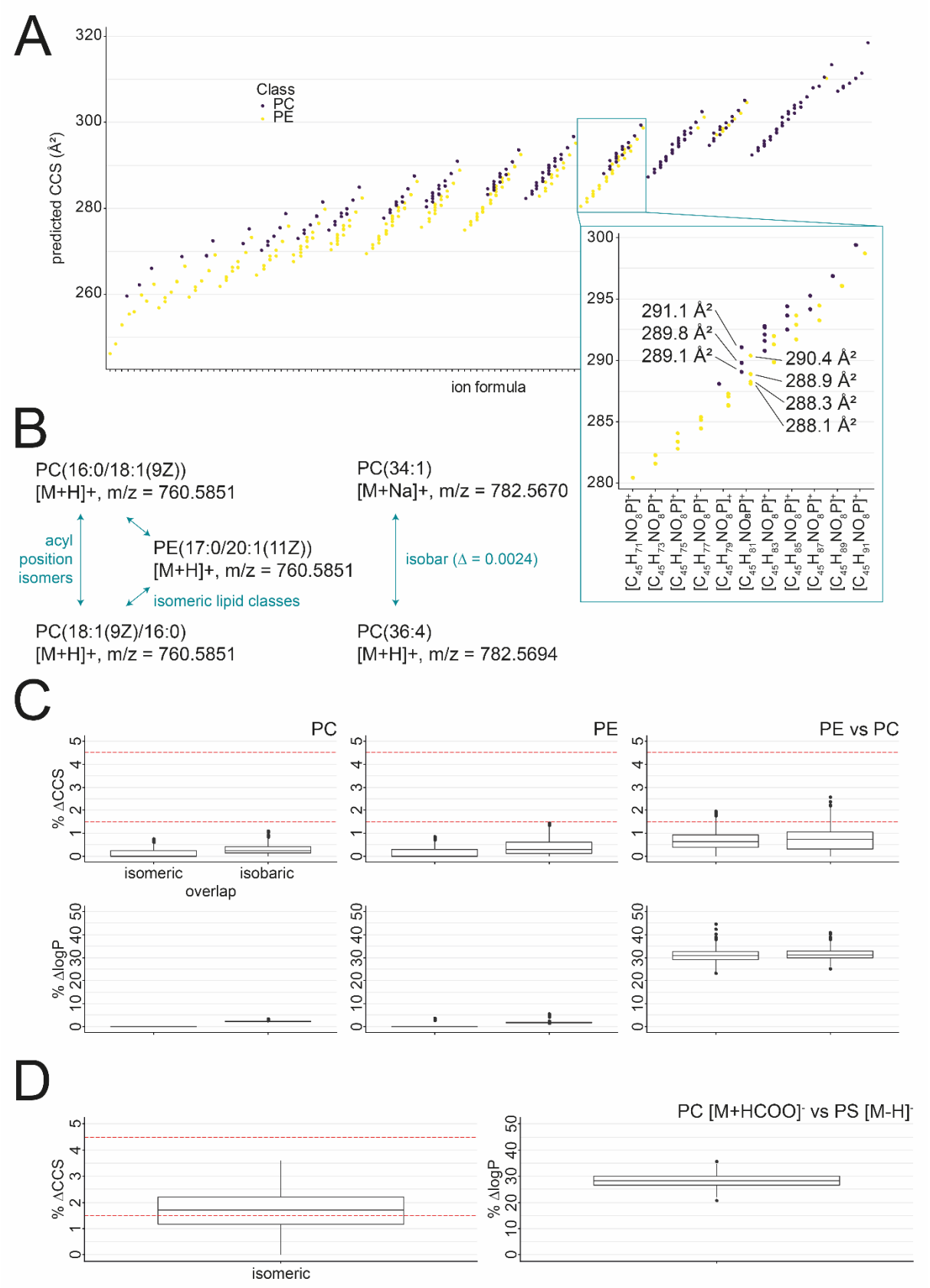


Figure 16 Evaluation of CCS values for the separation of isomeric and isobaric lipid species. (A) Comparison of CCS values for different ions obtained from $[M+H]^+$ of PCs and PEs. A certain overlap between the two classes in m/z and ion formula exist for C. elegans lipids. Within a lipid class, clear trendlines are visible, but separation within a single ion formula between the two lipid classes is often not possible due to the limited resolution. (B) Examples of isomeric and isobaric overlap for C. elegans lipids. (C) Differences in CCS were plotted against the

isomeric and isobaric overlap within and between PCs and PEs. Additionally, the difference in logP is used as proxy for the chromatographic separability of lipids. (D) Same as (C) but for negative mode and isomeric overlap between $[M+FA-H]^-$ adducts of PC and $[M-H]^-$ adduct of PS. Red dashed lines indicate differences of 1.5 and 4.5% as discussed by Causon et al. [387].

Recently, ion mobility has been combined with MALDI imaging. Although the additional separation dimension improves data quality, great care has to be taken when identifying lipids solely on *m/z* and CCS value. A prototype MALDT timsToF IMS has been used to analyze a whole-body mouse pup tissue has been reported. The instrument was able to show a beginning separation in the mobilogram between the isobaric PC 32:0 ionized as $[M+Na]^+$ and PC 34:3 $[M+H]^+$. Although peaks were separated, the resolution was not sufficient for baseline separation. Results indicate that at the current stage, resolution and accuracy are not high enough for unambiguous identification of lipids based on m/z and CCS values, but they provide a useful addition to RT, *m/z,* and MS^2 data. Especially trendlines, but also the additional separation can help in identification.

3.3.2.5. DTIMS-AllIons for identification of maradolipids.

The use of CCS values beyond dereplication by matching with known or predicted CCS values is possible. Trendlines and the additional separation dimension are potential additional filters for lipid identification and the reduction of false positive annotations. IM can be combined with data-independent fragmentation and can enhance the reconstruction of MS^2 spectra from this acquisition mode. One use case is maradolipids from *C. elegans* dauer larvae. To characterize the IM separation of maradolipids, $^{DT}CCS_{N2}$ values of authentic reference standards were determined. Maradolipid standards were infused in a 50/50 mixture of eluent A and B of the later employed chromatographic method. In positive ionization mode maradolipids are ionizing as $[M+NH_4]^+$ adducts during direct infusion as well as $[M+FA-H]^-$ adducts in negative mode. This is in agreement with Penkov *et al.*, who detected acetate adducts of maradolipids in negative ion mode [80]. Although $[M+Na]^+$ adducts were detected during chromatographic analysis, they were not detected in the direct infusion experiments. $^{DT}CCS_{N2}$ values of the maradolipid standards were determined using the stepped field method according to Stow *et al.,* similar to the phospholipids above [376]. Consistent with other lipid classes examined above, increasing chain length led to increased $^{DT}CCS_{N2}$.

In the next step, UHPLC-IM-QToFMS was performed using a single-field drift tube experiment. This allowed the collection of RT and $^{DT}CCS_{N2}$ in parallel. $^{DT}CCS_{N2}$ values from the single-field experiment were in good agreement with values derived from the multifield method (SI Table 6). Since lipids are based on defined building blocks, with increasing chain length in the fatty acid tails, this fact can be used to identify trends e.g. using KMD and RKMD and plotting it against *m/z*, RT, or CCS. They are calculated according to the following equations.

$$KM = exact\ mass \cdot \frac{14}{14.015650}$$

$$KMD = KM - nominal\ KM$$

$$RKMD = \frac{(experimental\ KMD - reference\ KMD)}{0.013399}$$

KMDs can be calculated for representative species of a lipid class. The KMDs for PCs, PEs, and PSs as $[M+H]^+$ are 0.7358239, 0.7357739, and 0.6765544. The KMD for maradolipids as $[M+H]^+$ is 0.6676465. To identify potential trends for investigations in natural samples, the KMD for CH_2 and RKMD for different lipid classes against the *m/z*, RT, and $^{DT}CCS_{N2}$. As expected, homologous series form horizontal lines at 0, -1, -2, etc. In contrast to glycerophospholipids, the maradolipids have no distinct sn1, or sn2 position since the 6 and 6’ positions on the trehalose are equal. Therefore, only single chromatographic peaks will be measured throughout the measurements, while for glycerophospholipids, two peaks might be found in the RT and IM dimension. However, as mentioned above, several isomers cannot be separated. Similar to PCs or PEs, maradolipids show a linear increase in $^{DT}CCS_{N2}$ with growing chain length. Slopes of trendlines for $^{DT}CCS_{N2}$ vs. *m/z* plots are slightly smaller for maradolipids compared to PCs and PEs (data not shown). In contrast to IM-MS alone, UHPLC-IM-Q-ToFMS was able to separate the isobaric structures Mar 16:0/16:0 and Mar 15:0/17:0. Trendlines along RT, *m/z*, RKMD, and $^{DT}CCS_{N2}$, therefore, can be combined to identify potential members of specific lipid classes. Potential isomeric and isobaric overlap might exist, depending on the induvial lipid class and ionization mode. In the case of

maradolipids, putative isomeric overlap within a five mDa window in negative ion mode with theoretical PE-Cers and SMs with a high number of hydroxyl groups was found using the LipidMaps search against CompDB [397]. Since such lipids are currently not known in *C. elegans* and not expected, collective information on the MS^1 level (*m/z*, RKMD, and $^{DT}CCS_{N2}$) allows identifying of putative maradolipid candidates in dauer larvae lipid extracts.

Still, this information is not sufficient to achieve higher identification levels, according to Schymanski *et al.,* and needs to be combined with MS^2. Fragmentation patterns of maradolipid standards were investigated using UHPLC-IM-Q-ToFMS/MS with a 4 Da isolation window and targeted fragmentation. First, fragmentation in negative mode was investigated. Fragmentation pathways of acetate adducts of maradolipids have been described by Papan *et al.* [154]. Upon fragmentation, first, the $[M-H]^-$ ion is formed from which the fatty acids are lost and can be detected as free acyl or as neutral losses. Subsequently, fragments with *m/z* 323.0984 and 305.0878 derived from trehalose are formed, corresponding to $[trehalose-H_2O-H]^-$ and $[trehalose-2\ H_2O-H]^-$.

Investigating the fragmentation of, in this case, $[M+FA-H]^-$ adducts, similar fragmentation was observed. Fragmentation data of Mar 14:0/14:0 and Mar 14:0/18:1 was closely examined, both representing a symmetrical and an unsymmetrical maradolipid. Similar to the fragmentation observed by Papan *et al.,* first, the fragmentation of the $[M+FA-H]^-$ to the $[M-H]^-$ ion was observed [154]. This further fragments by losing one of the two possible fatty acids attached at the 6- or 6'-position, which leads to $[M-R_1COOH]^-$ or $[M-R_2COOH]^-$ fragments. In the case of Mar 14:0/14:0, only one single fragment was and in the case of Mar 14:0/18:1, two fragments were observed. The corresponding $[R_1COO]^-$ and $[R_2COO]^-$ fragments were also observed. The fragments $[M-R_1COOH]^-$ and $[M-R_2COOH]^-$ were only observed upon fragmentation with 20 eV. 40 eV yielded the highest intensities of $[R_1COO]^-$ and $[R_2COO]^-$ fatty acyl fragments.

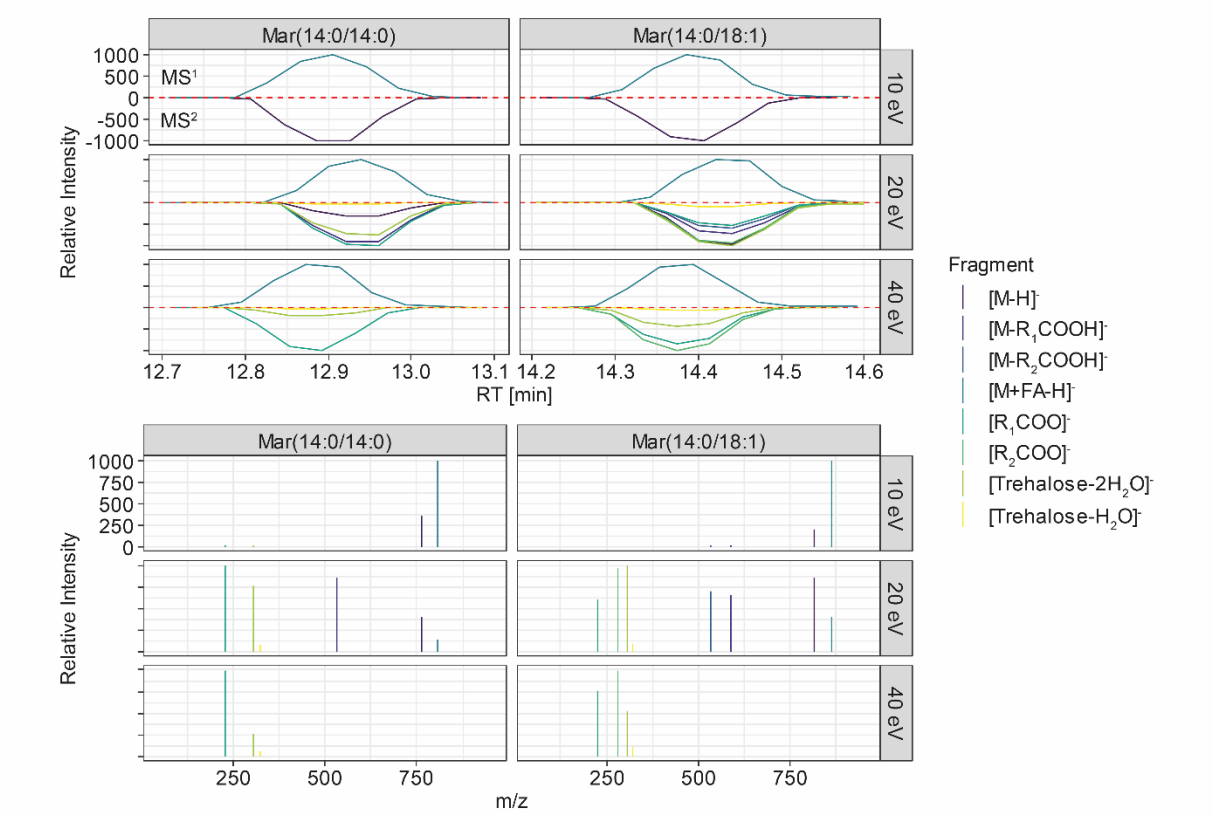


Figure 17: MS^2 spectra are reconstructed using correlation of drift time filtered EICs. Different fragmentation energies yield different amount of fragments with different intensities. Candidate peaks with a high correlation coefficient as well as an explainable fragment m/z are used for identification. The plots show the EICs for Mar 14:0/14:0 and Mar 14:0/18:1 as example.

In the used instrumentation, data-dependent fragmentation could not be combined with ion mobility separation. Therefore, data were collected using DIA fragmentation with alternating frames switching between low and high collision energy. Three different runs with either 10, 20, or 40 eV collision energy were produced. The aim was to investigate if UHPLC and IM-MS combined with DIA allow obtaining sufficient information for maradolipid identification. In this mode, RT, *m/z*, RKMD, and $^{DT}CCS_{N2}$ can be obtained from a single run without the need for multiple injections from the same sample. While DIA generally suffers from the loss of the precursor-fragment relationship, co-elution and similarity in drift times allow to filter the DIA MS^2 data and excluding false positive fragments, and re-establish such a relationship by correlation analysis. Based on the measurement of all available maradolipid standards, the elution profiles of precursors and fragments were investigated. EICs for the respective fragment *m/z* and drift region were generated and correlated against the EIC of the precursor in the respective retention time region. Generally, high correlation coefficients above 0.9,

indicating that, indeed, the correct fragments are assigned, were observed. Figure 17 shows examples for the two standards, Mar 14:0/14:0 and Mar 14:0/18:1. Investigation of positive ion mode fragmentation data showed that major fragments derived from $[M+NH_4]^+$ adducts are $[M-H_2O+H]^+$ as well as $[R_1CO]^+$ and $[R_2CO]^+$ of the two respective acyl groups. Since no additional information can be derived from combined positive and negative mode analysis, only negative mode data was further investigated.

Based on the obtained results, 20 eV was the most informative collision energy when performing non-targeted analysis and search for maradolipids since it yielded the most explainable fragments in a single collision energy. 40 eV yielded the highest intensity for fatty acyls and trehalose fragments and might be, therefore, the best choice for quantification workflows, e.g. using MRM or MRM-like approaches.

Next, to prove that the combination of RT, *m/z*, RKMD, $^{DT}CCS_{N2,}$ and DIA MS^2 can also identify maradolipids in biological extracts, *C. elegans* dauer larvae were generated from *daf-2(e1370)* mutants by growing them at 25 °C. Worms were harvested and extracted using a Bligh and Dyer extraction. Analysis of dauer larvae was performed by UHPLC-IM-Q-ToFMS using DIA fragmentation with either 10, 20, and 40 eV only in negative ionization mode. Detection of potential maradolipids in dauer larvae extracts was first performed with EICs for *m/z* 323.0972 and 305.0877 in the high collision energy frames generated (Figure 18). Coelution of these two *m/z* indicates the presence of potential maradolipids. Since 20 eV spectra contained the highest information content, they were investigated first. Indeed, coelution of the two *m/z* was observed in the range of 12 to 17 minutes, being in the same range where the standards are eluting. Interestingly, additional peaks for the *m/z* 323.0972 were observed in the range from 2.5 to 6.5 minutes but not for *m/z* 305.0867.

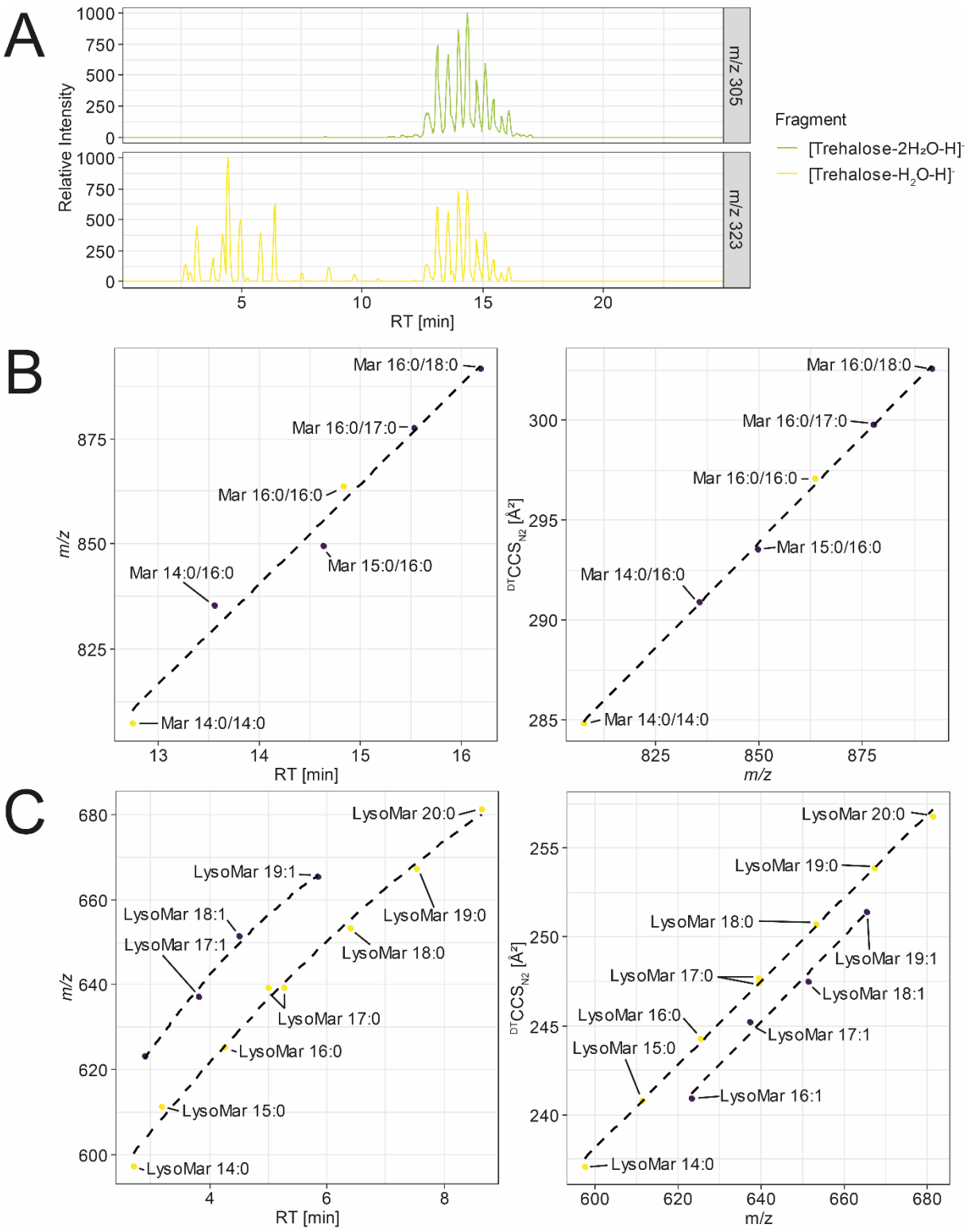


Figure 18: (A) Extracted ion chromatograms for m/z 305.0877 and 323.0972 in high collision energy frames of the UHPLC-IM-Q-ToFMS AllIons experiments. Coelution of both masses indicates presence of of maradolipids, while m/z 323.0972 alone indicates potential lysomaradolipids (B) RT and $^{DT}CCS_{N2}$ trendlines used for identification of Mar(14:0/16:0). Yellow points represent reference standards, while all others are derived from maradolipids detected in dauer larvae extracts. (C) RT and $^{DT}CCS_{N2}$ trendlines constructed for lysomaradolipids. For LysoMar(17:0), two peaks are visible. Two trendlines have been constructed, one for saturated and one for mono-unsaturated lysomaradolipids.

Indicating the presence of potential maradolipids in the dauer extract, non-targeted peak picking of lipid features was performed. In total, 1349 features were detected in all three replicates of dauer larvae lipid extract in negative ion mode. Based on the obtained *m/z*, RT, and $^{DT}CCS_{N2}$ library from reference standards, 31 phospholipids could be annotated. To identify potential maradolipids, the KM, KMD, and RKMD were calculated from the m/z values according to the equations above. Using an error of ± 0.1 for the RKMD the total list was narrowed down to 123 potential maradolipid candidates based on MS^1 information. The list was further condensed by filtering on the RT region of eluting maradolipid standards and compared against a computer-generated list of potential maradolipids using potential fatty acids present in maradolipids based on results from Penkov *et al.* [80]. Using MS^1 annotation to filter potential maradolipids, 33 candidates remained. Of these, ten could be matched with the used standards based on *m/z*, RT, and $^{DT}CCS_{N2}$ values as well as fragmentation pattern.

Investigating peaks that are putatively annotated as additional maradolipids, several interesting candidates were found. For example, *m/z* 835.5424 showed a small side peak in addition to the peak matched with the Mar 15:0/15:0 standard, which might represent an isobaric species with a different fatty acid composition. Investigating the DIA fragmentation data, it was putatively identified as Mar 14:0/16:0. To confirm further this putative identification, trends along RT and $^{DT}CCS_{N2}$ values were checked for maradolipids that contained 14:0 and 16:0 fatty acyl side chains. Mar 14:0/14:0 and Mar 16:0/16:0 have been measured as a reference standard. The putative Mar 14:0/16:0 falls between these standards in regard to RT and $^{DT}CCS_{N2}$. Although the Mar 16:0/16:0 standard deviation from the RT trendline was higher, trends along $^{DT}CCS_{N2}$ trend lines were fitting. Generally, a higher deviation of RT from standards was observed for maradolipids in *C. elegans* samples, but errors were generally below 2%, while the highest error for $^{DT}CCS_{N2}$ was 0.4%. Furthermore, $^{DT}CCS_{N2}$ trend lines showed good linear trends, while for RT, this was only the case for very limited examples and typically showed quadratic behavior. Combining all available information, the peak can be putatively to be Mar 14:0/16:0 based on DIA

fragmentation data, RT and $^{DT}CCS_{N2}$. Penkov *et al.* also detected Mar 14:0/16:0, and compared to Mar 15:0/15:0, it also showed lower levels [80].

While searching for potential maradolipids using DIA fragmentation, an additional region between 2.5 and 6.5 minutes showing the fragment *m/z* 323.0972 was identified. However, no corresponding fragment *m/z* 305.0877 was found. Therefore, it was hypothesized that the peaks in this area might represent lysomaradolipids. Papan *et al.* have identified lysomaradolipids using shotgun-based lipidomics analysis of lipid extracts from *C. elegans* dauer larvae [154]. The fragmentation pattern they have obtained shows strong similarities compared to the ones found in the present publication. Their proposed fragmentation matches the observation of the peaks eluting in this RT range. Using the obtained DIA fragmentation data, it was observed that a collision energy of 10 eV is more useful for the non-targeted search because the $[M+FA-H]^-$ and $[M-H]^-$ ions, as well as the $[trehalose-H_2O-H]^-$ fragments are present in the high collision energy data. For further structural elucidation, 20 eV collision energy was used since both the $[trehalose-H_2O-H]^-$ fragment, as well as fatty acyl fragments were visible, while 40 eV mostly produced the fatty acyl fragment. Using a similar filtering approach and putative annotation on the MS^1 level with masses of theoretical lysomaradolipids we identified a list of 12 potential candidates. Coelution of MS^1 *m/z*, the $[trehalose-H_2O-H]^-$ fragment, as well as specific fatty acyl fragments, were used for identification. Fragment *m/z* EICs were isolated for the specific drift time regions of the intact molecule, and peak correlation was performed. Figure 5A shows trendlines for lysomaradolipids.

Interestingly, for the *m/z* of LysoMar(17:0), two chromatographic peaks were found. While for the first and higher peak fragmentation data identified a fragment at *m/z* 269.2486 corresponding to a C17 acyl group, no fragmentation data confirming the putative ID was available for the second peak due to the low intensity of the precursor. However, while checking for co-elution with *m/z* 323.0972, perfect coelution could be observed for both peaks. *C. elegans* is able to produce mono-methyl-branched chain fatty acids on its own, and

most maradolipids contain a branched-chain fatty acid. One peak might be possible to represent a lysomaradolipid containing 15-methyl palmitic acid (FA 16:0(15Me)) and the other one heptadecanoic acid (FA 17:0). Both fatty acids have been detected in the analysis of total fatty acids, but heptadecanoic acid only in low amounts [292]. Investigating trendlines for both RT and CCS using odd-numbered LysoMar showed that both peaks match the trends between LysoMar 15:0 and LysoMar 19:0. However, if only higher peak eluting earlier is used, the fits increased. The $^{DT}CCS_{N2}$ value of the second peak is slightly higher (247.72 Å² compared to 247.44 Å²), which indicates a slightly larger structure. Since the two peaks showed good chromatographic separation, the logP value for both possibilities was calculated to measure hydrophobicity. The logP value of the hypothetical straight chain LysoMar 17:0 is 2.66, and the logP of the hypothetical iso-branched chain version is 2.50. This would fit with the trends seen based on $^{DT}CCS_{N2}$, indicating that the branched-chain version is eluting before the straight chain version. However, these identifications are only putative and need to be confirmed with authentic standards. Since no reference standards are currently available for lysomaradolipids these identifications cannot be further validated.

3.3.3. Conclusion

The usability of CCS for metabolite and lipid identification has been evaluated. While in the case of small molecules, CCS values are only of partial help, for lipids, they can be useful to reduce the number of false positive annotations. Investigations showed that for metabolites, CCS values of isomeric species are often too close to be separated in standard non-targeted experiments. Settings can be tuned towards the separation of specific isomeric species, which might be helpful for the quantification of individual species. No reference standards for *C. elegans*-specific metabolites are available; therefore, no measured CCS for e.g. ascarosides exist. However, prediction tools such as CCSBase or AllCCS can close this gap since predictions are typically accurate enough [350, 380]. However, RT, *m/z,* and MS^2 spectra are still of greater use for metabolite identification.

In the case of lipids, separation of isomeric and structures can be achieved for certain species, but for the large part, it is not possible. However, since lipids are made of defined building blocks with increasing chain length, they show a trend along a plot of CCS against *m/z*. These trendlines can be used to filter false positive annotations. While CCS alone might not be sufficient, in addition to *m/z,* it can help to remove potential false positive annotation. While CCS is often not enough to separate isomeric species, it can help in removing isobaric species. Both theoretical considerations and measurements proved this fact. Current developments like Structures for Lossless Ion Manipulations (SLIM) have shown that they have the potential for higher separation power. For example base line separation of the isomeric lipids PC 18:1(6Z)/18:1(6Z) and PC 18:1(9Z)/18:1(9Z) has been shown [398].

IM can be separated with DIA for lipid identification, as shown for the identification of maradolipids in *C. elegans* dauer larvae lipid extracts. In such cases, the IM dimension serves as an additional filter criterium for the reconstruction of MS^2 spectra from DIA data. Using a dedicated workflow, novel lysomaradolipids could be identified, including putative isomeric structures. IM for lipids works best when combined with chromatographic separation. However, the use of IMS leads to cleaner spectra, which can be used for better identifications. New developments like Parallel Accumulation-Serial Fragmentation (PASEF) enabled on TIMS instrumentation allows the fast collection of DDA spectra [399].

3.4. LipidFrag

Interpretation of the resulting MS^2 spectra, especially in high-throughput studies, is rather limited, and manual analysis of several hundred to thousands of MS/MS spectra is not feasible. To speed up identification, comparison against spectral reference databases is possible, but the lipid coverage in these databases is sparse. Lipid Maps currently contains only a few hundred low-resolution MS^2 spectra, while MassBank has 3,158 records on both low and high-resolution instruments covering 707 unique lipids [299]. Lipid identification is a major bottleneck in high-throughput lipidomics studies. However, tools for the analysis of lipid tandem MS spectra are rather limited. While the comparison against spectra in reference libraries is one of the preferred methods, these libraries are far from complete. *In silico* fragmentation has been suggested as a possible solution to analyze MS^2 spectra without the need for spectral reference databases [364]. LipidBlast is a spectral library that includes 212,516 *in silico*-generated tandem mass spectra covering 119,200 compounds from 26 lipid classes [400]. More recently, Greazy, an approach for the identification of phospholipids from MS/MS data, was presented, which includes the estimation of false discovery rates (FDR) [401]. The module LipidLama, integrated into Greazy, uses kernel density estimation to fit non-parametrized models to distinguish false and true lipid assignments. The cutoff score for a putative correct lipid assignment can then be defined by using a pre-defined FDR of e.g. 5 %.

LipidFrag is a workflow to improve the reliability of *in silico* MS^2 annotations of lipids. To achieve this, a bayesian classifier based on parametrized distributions and maximum-likelihood estimation was introduced to calculate a reliability score for a result to be a correct annotation among its lipid class, which is based on training data obtained from lipid standard materials and true positive manual identifications. This workflow consists of the annotation of precursor masses with possible lipid structures using MassTRIX [194, 195], followed by MetFrag batch processing of candidates retrieved via the putative neutral masses derived from ion species annotation results. The performance was evaluated using MS/MS spectra obtained previously with UPLC-UHR-ToF-MS/MS and data-dependent acquisition (DDA)

[68]. Results from this training allowed the development of the central new feature in LipidFrag, the classifiers to predict the probability of a reliable MetFrag annotation for an unknown lipid class. This is used to differentiate between good and poor identification results and to predict the underlying lipid main class of the precursor in high-throughput MS/MS experiments like in this case study performed with the lipid extract of *C. elegans*.

3.4.1. Material and Methods

3.4.1.1. Chemicals

HPLC-grade methyl-tert-butyl ether (MTBE), HPLC-grade chloroform ($CHCl_3$), LC-MS-grade methanol (MeOH), iso-propanol (iPrOH), acetonitrile (ACN), ammonium formate, and formic acid were obtained from Sigma-Aldrich (Sigma-Aldrich, Taufkirchen, Germany). Water was purified using a Merck Millipore Integral water purification system with a resistance of 18 MΩ and TOC < 5 ppb.

3.4.1.2. Lipid standard material preparation

Phosphatidylcholine preparation from chicken egg (840051P, Avanti Polar Lipids), Escherichia coli polar lipid extract (100600P, Avanti Polar Lipids), phosphatidyl serines from the porcine brain (840032P, Avanti Polar Lipids), ceramide from the porcine brain (860052P, Avanti Polar Lipids) and ceramide from chicken egg (860051P, Avanti Polar Lipids) were obtained from Avanti Polar Lipids (Otto Nordwald GmbH, Germany) and dissolved in MeOH at a concentration of 1 mg/mL. Additionally, L-alpha-Phosphatidylinositol sodium salt from Glycine max (P0639), Triglyceride mix (17811-AMP), 1,3-Dioleoyl-2-palmitoyl-glycerol (D1657), Glyceryl tritricosanoate (T1412), Glyceryl trioleate (T7140) and 1,2-Dilinoleoy-3-palmitoyl-rac-glycerol (D0301) were obtained from Sigma-Aldrich (Sigma-Aldrich, Taufkirchen, Germany) and dissolved in either MeOH, MTBE, $CHCl_3$ or solvent mixtures, depending on solubility. Different samples for analysis were prepared and diluted in H_2O/ACN/iPrOH (5/65/30, v/v/v) to 10 µg/mL for analysis.

3.4.1.3. Lipid extraction from C. elegans

Lipids were extracted from young adult *C. elegans* using a modified method from Matyash *et al.* [113], described in [68]. The worms were washed off the plates and their metabolism was quenched with 500 µL -20°C MeOH. Samples were flash-frozen in liquid nitrogen and stored at -80°C prior to extraction. Samples were then thawed on ice, 1.7 ml MTBE was added, and samples were vortexed vigorously. *C. elegans* were lysed for 30 minutes in an ice-cold ultrasonic bath, after which 420 µl of water was added, and samples were sonicated for further 15 minutes. Phases were separated by centrifugation at four °C and 14,000 rpm for 15 minutes. The upper organic phase was transferred to a 4 ml glass vial, and the remaining lower phase was re-extracted with an additional 650 µl MTBE for 15 minutes. After centrifugation, the organic layers were combined and evaporated in a SpeedVac vacuum concentrator at 30°C for 0.5-1h. The residue was redissolved in 500 µl H_2O/ACN/iPrOH (5/65/30, v/v/v).

3.4.1.4. UPLC-UHR-TOF-MS/MS-based lipid profiling

Lipid analysis was performed using two different methods. PCs were analyzed with the method published by Witting et al. [68]. Briefly, lipids were separated on Waters CORTECS UPLC C18 column (150 mm x 2.1 mm ID, 1.6 µm particle size) (Waters, Eschborn, Germany) using a linear gradient from eluent A (40% H_2O / 60% ACN + 10 mM ammonium formate / 0.1 % formic acid) to eluent B (10% ACN / 90% iPrOH + 10 mM ammonium formate / 0.1% formic acid) using a Waters Acquity UPLC (Waters, Eschborn, Germany) coupled to a Bruker maXis UHR-ToF-MS (Bruker Daltonics, Bremen, Germany). The following gradient was used: 68/32 at 0 min, 68/32 at 1.5 min, 3/97 at 21 min, 3/97 at 25 min, 68/32 at 25.1 min with a flow rate of 0.250 mL/min and a temperature of 40°C. The column was re-equilibrated for 2.5 minutes. Detection was carried out in positive and negative ionization modes with data-dependent acquisition with a scan rate of 5 Hz and selection of 2 precursors. Masses were excluded from DDA after three spectra and released from exclusion after 0.15 min. An absolute threshold of 1500 was used for selection.

3.4.1.5. MS data processing

MS data was imported to Genedata Expressionist for Mass Spectrometry 8.2 (Genedata, Basel, Switzerland) for internal re-calibration, retention time alignment, and peak picking. Files were exported to .xlsx format, and further data handling was carried out in MS Excel. Lipids were annotated with lipids from LipidMaps using the adducts $[M+NH4]^+$ and $[M+HCOO]^-$, as well as $[M+H]^+$, $[M+Na]^+$ and $[M-H]^-$ and an absolute error of 0.005 Da.

MS^2 spectra were exported from the calibrated and aligned chromatograms from Genedata Expressionist for MS 8.2 as a .mgf file. Only spectra associated with a detected feature were kept and converted to MetFrag batch files (available at http://msbi.ipb-halle.de/msbi/lipidfrag) using a custom Perl script. The neutral mass and formula for the batch file were obtained by annotation with lipids for all possible annotation results. Finally, spectra in batch files were de-isotoped using the CAMERA package with a custom R script [402].

3.4.1.6. Manual lipid identification

Manual lipid identification was performed using known lipid fragmentation pathways. Information from both ionization modes was combined and matched via identical retention times, where available. For phospholipids, fragments used for identification included head group fragments and their respective neutral loss, loss of fatty acid side chains, and their carboxylate fragment. In the case of triacylglycerols, neutral losses of fatty acid side chain as an ammonium salt and the respective fragments were used. Ceramide species were identified based on typical sphingolipid fragments, e.g. loss of N-bound fatty acid and sphingoid base fragments. Since the exact position and stereochemistry of double bonds cannot be deduced from these experiments, all possible isomers were reported as potential identification for further processing with LipidFrag.

3.4.1.7. LipidFrag identification

Batch query files were processed with the MetFrag command line tool (version 2.4 available at https://c-ruttkies.github.io/MetFrag/). Lipid Maps (LMSDFDownload18Mar14) was used as a structure database. Candidates were considered within 20 ppm of the theoretical mass,

and measured MS^2 peaks were matched against *in silico* fragments, generated with tree depth 3, with an error window of 0.01 Da + 15 ppm. The ion mode for the generated fragments were set according to the acquisition of the processed MS^2 peak list, and the minimum peak intensity was set to 1000 arbitrary units. The resulting ranked candidate lists were filtered by the first part of the molecules' InChIKey to eliminate stereoisomers and stored as CSV files, with the calculated MetFrag scores stored in the CSV columns. CSV files for MS^2 peak lists containing less than two informative MS/MS peaks were excluded from the evaluation. The score calculated by MetFrag was used to rank the known candidates of the standard spectra, and the pessimistic (worst case) ranking result when candidates, including the correct one, shared equal MetFrag scores were used. Hence all potential isomers, e.g. double bond positional isomers, which usually have identical MetFrag scores, are covered and reported.

The original MetFrag scoring function considers the bond dissociation energy (BDE) of bonds that are cleaved during the *in silico fragmentation*. As the cleavage of C-C bonds of the fatty acid chains is unlikely to occur under the given conditions in the mass spectrometer, the BDE of this bond type was set to the arbitrarily high value of 10e9, which effectively eliminates fragments generated by a C-C cleavage.

3.4.1.8. Lipid class-specific classifiers for reliability calculation

LipidFrag uses classifiers for the reliability calculation of the MetFrag result. The distribution of the MetFrag raw scores depends on both the query spectra and the lipid classes of the candidates. Generally, in metabolomics, this structural compound class classification is neither always obvious nor easy to obtain for small molecules, but for lipids, there is the structural categorization initiated by the International Lipid Classification and Nomenclature Committee (ILCNC), available on the LipidMaps website [207, 247]. This classification was used here to obtain well-defined ranges of MetFrag raw scores for particular lipid classes. Therefore, a training step was implemented to predict the reliability of MetFrag results based on the training of classifiers with MS^2 spectra of the lipids standard material for different lipid

subclasses. For this task, one classifier was created for each lipid subclass, where raw scores of correctly identified structures from the lipid standard materials served as foreground data. The same spectra were queried with deliberately wrong precursor candidates in the same mass range (up to 150 ppm), originating from the other lipid subclasses respectively, to obtain a decoy database and, subsequently, the MetFrag scores for the background data set. This approach was inspired by proteomics, where foreground and background training data are used to assign significance values to peptide identifications [403].

Gamma distributions were used to model the scores for the foreground and background data. The model parameters for the distributions were calculated by maximum-likelihood estimation on the fore- and background datasets. For each lipid class, a separate classifier was trained because the MetFrag scores exhibit large differences between the classes.

The following equation shows the calculation of the foreground class probabilities (FCP) of a MetFrag result with the bayesian approach, where P (score | Foreground, Θ) is the likelihood of the foreground model represented by a gamma distribution of the lipid subclass for the present score and *P (score | Background, Θ)* is the corresponding likelihood of the present score in the background model which is also represented by a gamma distribution. The estimated parameters of the distributions are represented by *Θ*.

$$FCP = \frac{p(score \vee Foreground, \theta)}{p(score|Foreground, \theta) + p(score \vee Background, \theta)}$$

For testing, 10-fold cross-validation was applied. FCPs of the lipid classes were used to calculate the true positive and false negative rates on the test dataset to determine a Receiver Operating Characteristic curve (ROC) and the Area under Curve (AUC) as a quality measure of the different classifiers.

3.4.1.9. Reliability of MetFrag results

After training, the classifiers were used to predict the reliability of MetFrag candidate identifications for the *C. elegans* MS/MS spectra, where the correct candidate is unknown.

Given a candidate list processed by MetFrag as SDF or CSV file, LipidFrag calculates the FCP for each candidate lipid in this result list by first selecting the appropriate classifier based on the candidate's Lipid Maps ID. The selected classifier, together with the calculated MetFrag raw score, is used to calculate the FCP value. Those results, where no candidate exceeds a defined FCP threshold (of e.g. 0.95), have to be treated as unreliable or not identified.

3.4.1.10. LipidBlast identification

For comparison, lipid annotations were performed using the LipidBlast *in silico* tandem MS library [18]. The provided LipidBlast fork (v2 Hiroshi Tsugawa fork) was downloaded and converted by the Lib2NIST tool (v1.0.4.38 (beta), options: "Include Synonyms": Yes, "MW from chem. formula": Yes, "MS/MS spectra only": Yes, "2008 MS Search compatible": Yes) to NIST format and used as a spectral library for LipidBlast annotation of all standards used for LipidFrag available in MGF format obtained from Genedata Expressionist for MS 8.2. The NIST MSPepSearchGUI (v0.91, options: defaults except for "Q-TOF": Yes, "Min. match factor": 100, "Presearch mode": Standard, "Load libraries in memory": No, "Max. number of output hits": 10, "Presearch mode": Standard, "Precursor ion tolerance": 0.02, "Fragment peak m/z tolerance": 0.02) was used to process input spectra in batch mode. LipidMaps identifiers provided for the correct identifications were mapped to common names annotated by LipidBlast for comparison with the LipidFrag annotations. The pessimistic rankings (among the top 10 reported candidates) were calculated based on the Rev-Dot (reverse dot) scores and compared with the LipidFrag results.

3.4.2. Results and Discussion

LipidFrag uses the result scores of a lipid candidate list retrieved from MetFrag, which performs *in silico* fragmentation of lipids. Then the matching classifier is selected based on the lipid subclass of a currently considered lipid candidate in the candidate list. Using the bayesian equation, LipidFrag calculates the posterior probability of the MetFrag score under the assumption to come from the foreground distribution of the selected bayesian classifier.

This probability value can then be used as a prediction of the lipid class of the regarded MS/MS spectrum and, secondly, as a measure of the reliability of the current MetFrag lipid annotation to filter out false positive lipid assignments.

3.4.2.1. Analysis of lipid standard materials

For the positive ionization mode spectra, classifiers were built for the following lipid subclasses: PC (LMGP0101), PE (LMGP0201), PS (LMGP0301), PI (LMGP0601), Cer (LMSP0201/ LMSP0202) and TG (LMGL0301). As the scores for the Cer species (LMSP02) show a bimodal distribution in positive ion mode, two separate classifiers were trained for the available ceramide subclasses (LMSP0201 and LMSP0202) for the foreground data. Compared to a single classifier for the whole Ceramide main class, this captures the multimodal score ranges of the lipid subclasses in a better way (SI Figure 1). For the negative ionization mode spectra, the lipid subclasses: PC (LMGP0101), PE (LMGP0201), PS (LMGP0301), PG (LMGP0401), PI (LMGP0601), and Cer (LMSP0201/ LMSP0202) were used for training. As candidates for the LMGP0101 subclass show similar MetFrag scores on LMGP0201 subclass MS^2 spectra, a combined classifier was trained. This resulted in six different classifiers for positive and five for negative ionization mode. With these classifiers, LipidFrag is able to cover already over one-third of the lipid species in the Lipid Maps database.

The classifiers were extensively cross-validated on the lipid standards spectra to generate receiver operating characteristic (ROC) curves and the corresponding area under the curve (AUC) values as a measure of identification performance. For clarity, results are grouped into three lipid types: ceramides, glycerophospholipids, and glycerolipids, and presented separately in the following paragraphs. Mean ranks shown in Table 18 and Table 19 are calculated with and without an FCP threshold to highlight the performance using the LipidFrag classifiers. To reduce the false negative rate an FCP threshold of 0.6 was set within LipidFrag. With this value, the number of false positive assignments could be reduced from 91% to 57% for positive ion mode and from 93% to 27% for negative mode.

Table 18: Results of the MetFrag identification and the classifier testing. The mean values of the FCPs retrieved from the cross-validation for the foreground (FCP+, higher is better) and the background (FCP-, lower is better) scores are shown. An AUC of 1.0 represents the best possible classification result for the corresponding lipid main/sub class. Additionally, the mean rank of the correct candidate (Rank) using MetFrag and LipidFrag with a FCP threshold of 0.6 together with the discarded proportion of false positives (FP-Rate) and the mean number of candidates retrieved (Cand) are given.

Metric	LMGL0301 (TG)	LMGP0101, LMGP0201 (PC, PE)	LMGP0201 (PE)	LMGP0301 (PS)	LMGP0401 (PG)	LMGP0601 (PI)	LMSP0201, LMSP0202 (Cer)
FCP+	---	0.871	---	0.979	0.888	0.834	0.817
FCP-	---	0.098	---	0.009	0.164	0.154	0.236
AUC	---	0.979	---	1.0	0.901	0.961	0.931
Mean Rank	---	2.2	---	1.8	1.8	2.6	1.3
Mean Rank (FCP >= 0.6)	---	2.3 (68%)	---	1.8 (55%)	1.8 (94%)	2.4 (78%)	1.2 (63%)
Cand	---	31.3	---	15.8	15.3	14.6	2.3

Table 19. Results of the MetFrag identification and the classifier testing. The mean values of the FCPs retrieved from the cross-validation for the foreground (FCP+, higher is better) and the background (FCP-, lower is better) scores are shown. An AUC of 1.0 represents the best possible classification result for the corresponding lipid main/sub class. Additionally, the mean rank of the correct candidate (Rank) using MetFrag and LipidFrag with a FCP threshold of 0.6 together with the discarded proportion of false positives (FP-Rate) and the mean number of candidates retrieved (Cand) are given.

Metric	LMGL0301 (TG)	LMGP0101 (PC)	LMGP0201 (PE)	LMGP0301 (PS)	LMGP0401 (PG)	LMGP0601 (PI)	LMSP0201, LMSP0202 (Cer)
FCP+	1.000	0.551	0.994	0.969	---	1.000	0.908
FCP-	0.000	0.442	0.000	0.000	---	0.000	0.095
AUC	1.0	0.799	1.0	1.0	---	1.0	0.935

Metric	LMGL0301 (TG)	LMGP0101 (PC)	LMGP0201 (PE)	LMGP0301 (PS)	LMGP0401 (PG)	LMGP0601 (PI)	LMSP0201, LMSP0202 (Cer)
Mean Rank	3.1	5.8	1.7	1.9	---	1.0	1.17
Mean Rank (FCP >= 0.6) (FP-Rate)	1.7 (16%)	3.0 (45%)	1.7 (49%)	1.9 (9%)	---	1.0 (100%)	1.0 (68%)
Cand	33.9	26.5	26.5	14.9	---	15.3	2.2

3.4.2.2. Ceramides

Ceramides have quite distinct molecular formulas compared to other lipid classes (i.e. glycerophospholipids); therefore, the overlap with other classes and the number of potential candidates is low. Major differences between different ceramide species are the length of the sphingoid base, the number of hydroxyl groups in the sphingoid base, the length of the N-linked fatty acid, and the total number of double bonds. The fragmentation of ceramides has been studied extensively by Hsu et al., focusing mainly on the $[M-H]^-$ ions, whereas here ceramides were observed predominantly as $[M+HCOO]^-$ adducts in negative ionization mode [404]. Both positive and negative ionization modes were used to characterize the ceramides. In total, 17 ceramides were identified manually from obtained MS^2, with 11 found in both ion modes, 2 in negative and 4 in positive ion mode only.

LipidFrag shows the best ceramide results in positive ionization mode, indicated by the AUCs of 0.935 for the two ceramide classes (LMSP0201 /LMSP0202). In negative ionization mode, the AUC is also good, with a value of 0.931. The mean rank of the correct solution is 1.17 in positive and 1.3 in negative ionization mode, which is also due to the low number of potential candidates.

3.4.2.3. Glycerophospholipids

Different classes of glycerophospholipids were subjected to fragmentation, including PC (LMGP0101), PE (LMGP0201), PS (LMGP0301), PG (LMGP0401), and PI (LMGP0601). The molecular formulas of PCs and PEs overlap considerably, which can lead to ambiguous results if only the accurate mass of the precursor is used for the annotation with potential structures. Ekroos *et al.* studied the use of fragmentation and fatty acid scanning using an ion trap MS for elucidation of the fatty acid composition of PCs [405]. Fragmentation is very class and ion mode specific, e.g. PCs yield mainly m/z 184.07 as the head group fragment in positive ionization mode, whereas in negative ionization mode, fragments originating from $[M+HCOO]^-$ adducts provide information about fatty acid composition. Diagnostic fragments indicating fatty acid composition were only detected for very highly abundant species in positive ion mode. Several studies have shown that the carboxylate anion from the sn2 fatty acid is up to three times higher compared to sn1 [406]. PEs in contrast, show mainly the diacylglycerol fragment derived from the neutral loss of the head group in positive ionization mode and side chain fragments of very low intensity (usually below 2%). Therefore, MS^3 of the diacylglycerol fragment is needed for side chain identification in positive ion mode. In negative ion mode, fragmentation of PE species yields carboxylate anions from sn1 and sn2 fatty acids similar to PCs. Most of the glycerophospholipids show very good identification results with LipidFrag. This is indicated with the mean rank values 2.24, 1.8, 1.8 and 2.6 for the available PC/PE (LMGP0101 /LMGP0201), PS (LMGP0301), PG (LMGP0401) and PI (LMGP0601) species in negative ion mode. The AUCs of 0.979, 1.0, 0.901, and 0.961 also show excellent classification results. In positive ionization mode, the PE (LMGP0201), PS (LMGP0301), and PI (LMGP0601) species show similar results with mean ranks of 1.7, 1.9, and 1.0 and the AUCs of 1.0. Though the PC (LMGP0101) species show a similar performance with a mean rank of 1.69 when using an FCP filter with a threshold of 0.6, the filter sorted out 58 of the 71 spectra caused by the limited fragmentation, which also indicated by a lower AUC of 0.799.

3.4.2.4. Glycerolipids

Glycerolipids (LMGL0301) were mainly detected in positive ionization mode as $[M+NH4]^+$ adducts. From this adduct, typical fragmentation is the neutral loss of fatty acid side chains plus ammonia, yielding a diacylglycerol-like fragment [407]. This loss can occur for all side chains and lead to a pattern that allows the identification of composition but rarely provides sufficient information to determine the position of fatty acids in the intact lipid. Five different TG standards were employed as training data, showing previously known fragmentation pathways. These five compounds had different fatty acid compositions and, therefore different retention times. However, in *C. elegans* samples, many possible isomers and isobars are co-eluting with many different fatty acid combinations that can be deduced from fragmentation data. The TG species are observable in positive ionization mode, and the related classifier shows a good result with an AUC of 1.0. However, the mean rank indicates a lower performance for the identification results with 3.1, as the typical loss of a fatty acid side chain during fragmentation is not only explained by the correct candidate but also by structurally very similar TG species. The fragment peaks of these types of losses seem to be very specific for the main lipid class, indicated by the high AUC, but this does not help to distinguish between different TG lipids sharing the same molecular formula.

3.4.2.5. Handling of mixed spectra

One potential problem, not only for LipidFrag, is non-pure spectra arising from the co-isolation of co-eluting isomeric/isobaric lipids during the MS measurement. In order to test how well LipidFrag can deal with this, we created such spectra *in silico* using measured spectra as a template. Overlap especially occurs for glycerolipids in the later elution range of the chromatogram but might also occur for other lipids. Although the UPLC method used can separate major isobars of the glycerophospholipids [23], overlap might also occur with major interference most likely coming from isomers/isobars within the same lipid class. Interference from different lipid classes having the same molecular formula can be neglected because polarity, and hence retention time is very different (e.g. PE 18:0/20:2 has a logP of 13.12, whereas the isobaric PC 18:0/17:0 has a logP of 11.47). Measured lipid MS^2 served as the

target, and interfering MS/MS peaks at the intensity ratios of 10:1, 2:1, and 1:1 were added, and the MetFrag raw score of the true candidate was evaluated. Mixtures included binary, ternary, and even quaternary mixes of isobaric lipids. Results indicate that mixtures with an equal amount of target and interference cause a drop in the score and rank of the true candidate, as expected. The target substances still rank in the upper quarter. Results from one particular example in *C. elegans* samples having two isomeric PC species in one MS^2 spectrum are discussed in a later section.

3.4.2.6. Analysis of publicly available MSMS spectra

To test the performance of the LipidFrag approach on an independent second dataset, we used 415 negative ion mode lipid MS/MS spectra retrieved from MassBank, where a Lipid Maps ID was available for the correct candidate. Although these spectra were measured on a different instrument with higher mass error than the data used for classifier training, they served as an additional validation of the workflow. Altogether, the spectra were annotated by the submitters to be from ten different subclasses (LMGL0301, LMGP0101, LMGP0102, LMGP0103, LMGP0105, LMGP0201, LMGP0202, LMGP0203, LMGP0601, and LMSP0301). Table 20 shows the ranking results obtained from LipidFrag. The mean ranks within the lipid subclasses were 4.4, 6.0, 2.9, 3.9, 2.3, 2.8, 1.0, 1.0, 2.0, and 1.8, respectively. Only two classifiers were available for the spectra originating from PC/PE (LMGP0101 /LMGP0201) and PI (LMGP0601) species. For the 180 MS/MS spectra, 157 have been identified with the correct lipid subclass based on the foreground class probability (FCP), which is a true positive rate of ~87% for the low-resolution spectra where a classifier was available. The LMGP0601 classifier calculated a subclass FCP which reached this threshold for all cases (9) and the LMGP0101/LMGP0201 classifier for 148 out of the 171 cases.

3.4.2.7. Comparison with LipidBlast annotations

The results of LipidBlast compared with the mean ranks of LipidFrag are shown in Table 21 and Table 22 for the negative and positive ionization modes, respectively. The values indicate that results are comparable between both software tools. Nevertheless, there are

slight deviations for some lipid classes, and LipidFrag usually annotates more spectra (FCP threshold 0.6) for both ionization modes.

Table 20: LipidFrag rankings on the 415 MassBank spectra

Lipid sub class	Mean rank	Median rank	Mean candidates	Median candidates	Number MS/MS
LMGL0301 (TG)	4.4	2.0	15.0	15.0	7
LMGP0101 (PC)	6.0	3.5	22.0	23.0	118
LMGP0102 (PC)	2.9	3.0	9,2	8.0	36
LMGP0103 (PC)	3.9	2.5	15.7	14.0	18
LMGP0105 (PC)	2.3	2.0	4.2	4.0	30
LMGP0201 (PE)	2.8	2.0	17.2	19.0	53
LMGP0202 (PE)	1.0	1.0	7.0	7.0	12
LMGP0203 (PE)	1.0	1.0	12.5	14.5	24
LMGP0601 (PI)	2.0	2.0	11.3	11.0	9
LMSP0301 (SM)	1.8	1.0	18.9	14.0	108
All	3.3	2.0	16.6	16.0	415

In positive ionization mode, on average LipidFrag could annotate 69 and LipidBlast 49.7 spectra across all lipid classes. Considering the mean ranks, LipidBlast showed better results for TG (LMGL0301) (1.0 to 3.1) species. No results were annotated for PI spectra, as the predictions are missing in the current spectral database mirror of LipidBlast. Developers of

LipidBlast indicated that predictions are in progress for several missing lipid classes and will be added to the library in the near future. LipidFrag showed better mean ranks for PE (LMGP0201) (1.7 to 1.8) and PS (LMGP0301) (1.9 to 7.8) species. Equal mean ranks for both software tools could be assigned to the Ceramide classes (LMSP0201 and LMSP0202) with a value of 1.0. Both software tools filtered out a large proportion of the PC spectra (LipidFrag: 58 spectra, LipidBlast: 62 spectra) as this lipid class shows sparse fragmentation in positive ionization mode resulting in less informative MS^2 spectra.

For negative ionization mode, LipidFrag and LipidBlast could annotate an almost equal number of MS/MS spectra with mean values of 81 and 83.6 across all lipid classes. LipidBlast performed slightly better for the Cer (LMSP0201) (1.0 to 1.6) and the PI (LMGP0601) species, whereas LipidFrag showed better mean ranks for PC/PE (LMGP0101 /LMGP0201) (2 to 2.3) and PG (LMGP0401) (1.0 to 1.8) species.

Table 21: Comparison against LipidBlast (negative ionization mode)

Mean Rank	TG (LMGL0301)	PC/PE (LMGP0101/LMGP0201)	PC (LMGP0101)	PE (LMGP0201)	PS (LMGP0301)	PG (LMGP0401)	PI (LMGP0601)	Cer (LMSP02/LMSP0202)
LipidFrag	---	2.3 (112)	---	---	1.8 (35)	1.8 (41)	2.4 (62)	1.2 (155)
LipidBlast	---	**1.2 (116)**	---	---	**1.0 (34)**	**1.0 (40)**	**2.3 (70)**	**1.0 (158)**

Table 22: Comparison against LipidBlast positive ionization mode

Mean Rank	TG (LMGL03)	PC/PE (LMGP0101/LMGP0201)	PC (LMGP01)	PE (LMGP02)	PS (LMGP03)	PG (LMGP04)	PI (LMGP06)	Cer (LMSP0201/LMSP0202)
LipidFrag	3.1 (25)	---	1.7 (13)	**1.7 (88)**	**1.9 (50)**	---	**1.0 (82)**	**1.0 (156)**
LipidBlast	**1.0 (13)**	---	**1.0 (9)**	1.8 (75)	7.8 (43)	---	NA (0)	**1.0 (158)**

3.4.2.8. *Analysis of C. elegans samples*

To demonstrate the applicability to biological data, lipids extracted from *C. elegans* were used, representing a real challenge for LipidFrag. Several lipid classes are present in the worm, and different fatty acid combinations, including odd-numbered side chains, are possible in glycerol- and glycerophospholipids. Shotgun lipidomics was applied for the analysis of a novel class of lipids only present in dauer larvae [80]. LipidFrag was then applied to MS^2 spectra obtained from *C. elegans* lipid extracts. Table 23 gives an overview of detected lipid features in positive and negative ion mode runs. Altogether 1,518 MS/MS spectra acquired in negative, and 2,355 MS/MS spectra in positive ion mode were processed. Results with a foreground class probability (FCP) of ≥ 0.95 can be found across the whole intensity range, although higher intensities seem to lead to better results in positive ion mode. More important than precursor intensity is to detect diagnostic fragments, especially in negative ion mode, where fatty acyl side chains can be directly detected. Good results in this mode were also obtained for most of the middle-intensity range. Table 23 gives an overview of the number of detected lipid features, their corresponding MS/MS information, and LipidFrag results.

Table 23: Number of detected features in C. elegans samples with corresponding MS/MS and LipidFrag results

Ionization mode	No. of detected features	With accurate mass annotation	With MS/MS	Manually identified in standards	Reliable LipidFrag (FCP cut-off)
Positive	1655	1297	685	65	108 (0.7) 106 (0.8) 100 (0.9) 98 (0.95)
Negative	505	358	228	52	45 (0.7) 43 (0.8) 43 (0.9) 40 (0.95)

For the 3,873 (1,518 + 2,355) *C. elegans* spectra used, the MetFrag *in silico* fragmentation and scoring took altogether ~31 hours (user+system time) on a single core CPU, i.e. 30 seconds per spectrum. Using the calculated classifiers, which are based on the standard lipid

spectra, the FCP calculation for all 3,873 *C. elegans* spectra took less than 10 minutes or 0.15 seconds per spectrum.

For the positive ionization mode, LipidFrag detected 452 spectra as TG (LMGL0301), and 69 as PE (LMGP0201). Additional 3 PE and 1 PC (LMGP0101) species were added by decreasing the FCP threshold to 0.9. In negative ion mode, LipidFrag found 206 spectra with PC/PE (LMGP0101/LMGP0201lipid subclass annotations having an FCP ≥ 0.95. With a lower FCP threshold of 0.9, additional 47 PC/ PE species were annotated. Irrespective of the ionization mode, over 22% of the LipidFrag results have an FCP ≥ 0.75.

One example is the spectrum of PE 18:0/20:5. The most prominent peaks show the corresponding fatty acids, with higher intensities for 20:5 bound at the sn2 position. A further diagnostic fragment $[M\text{-}sn2\text{-}H]^-$ at *m/z* 480.3096 is detected, and with lower intensities also the $[M\text{-}sn1\text{-}H]^-$ at *m/z* 498.2626 ions. Precursor mass, these four peaks, and their respective ratios allow manual identification as PE 18:0/20:5. Furthermore, the head group was detected as a fragment with a fragment containing the head group and the glycerol backbone. MetFrag was able to explain 8 of 9 fragments for identification. Additionally, small fragments derived from 20:5 were found. LipidFrag calculated an FCP of 0.91 for the result being a PE. The fatty acid positional isomer showed a similar score and probability. Because the scoring does not take any intensity ratios into account, both isomers obtain the same score. At this point, manual interpretation of intensities is required to determine which annotation is correct. The isobaric PC 15:0/20:5 was ranked third, with a similar MetFrag raw score (103.49998 for the correct PE and 96.64367 for the PC) but a lower FCP of 0.86 and only 7 of 9 peaks correctly explained. The number of explained peaks was used as an additional metric for correct identification in case scores and probabilities are similar. Lipids in the used biological samples subjected to fragmentation by DDA were almost exclusively from the class of glycerophospholipids, di- and triacylglycerols. Using *C. elegans* lipid extracts, it was shown that the developed approach could be applied to biological samples. Coverage of features with one or more associated MS/MS spectra has to be improved, e.g.

using pseudo-targeted methods, data-independent approaches, and spectra reconstruction or improved DDA [408-410].

3.4.3. Conclusion

Although the number of tools for the automatic identification of lipids is increasing, most research still performs manual inspection of MS^2 spectra or automated comparison against rather small reference libraries. *In silico* fragmentation offers an elegant, automatic way to tentatively identify metabolites and lipids if no standard is available by reducing the number of possible candidates or even proposing just a single reliable match. A workflow was developed and validated for analysis of lipid MS^2 spectra derived from data-dependent acquisition on a UPLC-UHR-TOF-MS/MS system. This workflow is based on annotation of potential lipids to the precursor mass, isotope clean-up of MS^2 spectra, and identification using the *in-silico* fragmentation tool MetFrag in combination with a novel reliability calculation based on bayesian classifiers. Lipid standard materials were used for validation purposes, and the *in-silico* analysis was compared against manual identification.

Cross-validation of the obtained results showed that the true, correct identification could be easily separated from background spectra for most cases. Scores of correctly identified lipids and deliberately wrong candidates as decoys were used to generate fore- and background datasets to calculate the FCP giving reliability of a result of an unknown to be correct. Using lipid standard materials, the good performance of LipidFrag was shown, with high relative rankings of the correct candidate, high probabilities, and high AUC values obtained from the cross-validation. Furthermore, a comparison with LipidBlast, one of the most utilized tools for lipid spectra prediction, showed comparable results for both tools, with the main difference that the LipidFrag approach needs an initial training step for its classifiers but no *ab initio* information on fragmentation compared to LipidBlast. The workflow was applied to a lipid extract of *C. elegans*. From the obtained spectra, about 20% had high foreground class probabilities of ≥ 0.9. Higher identification rates could be achieved in future investigations by measuring more lipid standards from different classes to train more classifiers. However,

even with only 11 classifiers, the application of LipidFrag to MS^2 spectra derived from lipid extracts from *C. elegans* was successful and showed the advantage of this workflow.

An advantage here is that MetFrag does not rely on previously known fragmentation pathways and is therefore also applicable to novel lipid classes, currently not present in databases. In this case, candidate structures can be scored by generating potential structures, e.g. using theoretical lipids from LipidHome or even structures from a molecular structure generator like MOLGEN as an input database [411]. One particular example is the sphingolipids of *C. elegans*, containing a branched-chain C17 sphingoid base, which are not present in LipidMaps.

For the results retrieved from the *C. elegans* data, a comparison of the LipidFrag annotation with high probabilities and the manual identification for randomly-selected spectra showed excellent agreement with most of the peaks correctly explained by the *in silico* fragmentation. For application to complete lipidomics studies, the results from LipidFrag can serve as first filtering and interpretation for further manual investigation, especially for potential marker peaks. A major limitation is the co-elution of isomeric species leading to mixed MS^2 spectra. Although the chromatographic method is able to resolve several isomeric, not all of them can be resolved, especially for lipids like TGs, where several isomers exist. Where identified spectra as training data are available, e.g. through authentic standards, LipidFrag can help in high-throughput identification. With the standard MS setups, as employed in this study, lipid class and fatty acid composition can be deduced. Our selected example with the PE 18:0/20:5 species from the biological dataset showed that the MetFrag score alone could not distinguish ambiguous results. Here, the wrong candidate had a similar score to the correct one, but their FCPs were significantly different and enhanced the annotation confidence. Manual interpretation of obtained data often allows to additionally identify fatty acid position based on intensity ratios of fatty acid fragments, which is not possible with MetFrag.

The resulting output could be simplified by the lipid annotation scheme proposed by Liebisch *et al.,* which combines different lipid isomers under a common identifier [207, 247]. For mass

spectrometry using UHR-ToF-MS, the fatty acid scan level, and fatty acid positional isomer are relevant. The former represents lipid identification of the fatty acid composition, but their position is not determined. This level is well suited for LipidFrag identification. For example, all isomeric results can be collapsed under a common identifier, which would be easier to interpret. Unfortunately, the Liebisch annotation is currently not widespread in structural databases. LipidHome is an *in silico* database using this nomenclature, whereas no structural representation of the chemical structure is available, which would be needed for MetFrag [412]. Currently, no chemoinformatics representation of molecules exists to encode ambiguity in the position of double bonds.

3.5. wormLipidBlastR

For lipids known to occur in specific species, MS^2 spectra from reference standards can be used to be matched against the measured spectra, if available. This allows a fast first-line identification of previously detected lipids. However, the number of commercially available lipid standards is low (with only a few tens to hundreds). Even specialized companies offer only a limited number of reference standards, which cannot cover the entire range. *In silico* tools, such as LipidFrag described above, can help to close this gap. However, one major disadvantage of LipidFrag is that it relies on lipid structures present in structural databases, such as LipidMaps [205, 397]. Typically, only fully structurally elucidated lipids are added to such databases. Since often lipid characterization is not performed until full structure, certain identified lipids will never be added to such a database. However, such partial identifications can be used and might be of further interest in other datasets. Fragmentation patterns of lipids are often known and can be used to annotate lipids.

While in shotgun and targeted lipidomics, known fragmentation pattern and transitions of specific building blocks are used, non-targeted lipid profiling generates a large number of MS^2 spectra. While metabolites are highly diverse in their structure, which complicates the prediction of their fragmentation pattern, lipids in contrast, are made of defined building blocks, and their fragmentation follows a certain set of rules. The fragmentation of different lipid classes and species has been extensively studied. For example, Hsu and Turk published a collection of papers studying the fragmentation of different lipid classes and different ionization modes and adducts [406, 413, 414]. They collected extensive knowledge on fragmentation pathways. Once established, these rules can be used to generate *in silico* MS^2 spectra for unknown lipids. However, at the current stage, only the *m/z* value can be predicted accurately. For precise matching of library spectra and measured spectra, the intensity also plays an important role. However, the intensities of the different fragments are heavily influenced by the geometry of the collision cell, the type of collision gas, and the collision energy. Therefore, spectra generated on the same machine under identical analytical conditions represent the best option for spectral matching. In the case of lipids,

fragment spectra from measured lipid species can serve as a template to predict the spectra of different species. This is performed by LipidBlast [415, 416]. Identification of lipids relies on one side on known fragmentation pathways and building blocks and on the other on available reference spectra to derive data. LipidBlast is based on measurements of lipids from different LC-MS platforms together with rules of lipid fragmentation. In the first iteration, in total, 212516 spectra were generated from 119200 compounds from 26 lipid classes. Several add-on libraries have been generated, e.g. for glucuronosyldiacylglycerol (GlcADG) or FAHFAs [416, 417].

These libraries are either generated with a rather broad coverage or focus on specific lipid classes. However, it is unlikely that a specific organism will synthesize all possible lipid structures; this approach might lead to a too large search space. *C. elegans* is known to produce several lipids that are different from mammalian lipids (e.g. sphingolipids). Therefore, a subbranch named WormLipidBlast was developed. This library is based on lipids that are predicted from lipid reactions in WormJam, as discussed in Chapter 3. In contrast to LipidBlast, WormLipidBlast has been programmed in R instead of Excel, which allows more flexibility in its use. The core of this implementation is parsers for lipid shorthand notations that allow the generation of specific fragments based on lipid fragmentation rules and use template spectra to reconstruct the ion intensities. The basic principle is similar to the original LipidBlast, but it can handle different lipids in a simpler manner compared to the large Excel files. However, if required, new databases can be generated on the fly during data analysis.

3.5.1. Material and Methods

3.5.1.1. Chemicals

Methanol (MeOH), 2-Propanol (iPrOH), and acetonitrile (ACN) were of LC-MS grade (Sigma-Aldrich, Taufkirchen, Germany). All other solvents and chemicals were of the highest available purity, usually analytical grade. Water was purified on Merck Millipore Integral 3 water purification system with TOC < 3 ppb, 18 MOhm. Reference standards were

purchased from different chemical vendors (e.g. Sigma-Aldrich, Merck, Cayman etc) and dissolved in an appropriate solvent. Aliquots of stock solutions were stored at -20°C until further analysis. Lipid reference standards were purchased from Avanti Polar lipids. All used lipids are summarized in SI Table 3.

3.5.1.2. Lipid Template measurements on LC-MS

Two different LC-MS systems have been employed. First, a Waters Acquity UPLC was coupled to a Bruker maXis UHR-ToF-MS, and second, an Agilent 1290 Infinity II UHPLC was coupled to an Agilent 6560 IMS-Q-ToF-MS. Briefly, lipids were separated on Waters CORTECS UPLC C18 column (150 mm x 2.1 mm ID, 1.6 µm particle size) using a linear gradient from eluent A (40% H_2O / 60% ACN + 10 mM ammonium formate / 0.1 % formic acid) to eluent B (10% ACN / 90% iPrOH + 10 mM ammonium formate / 0.1% formic acid). The following gradient was used: 68/32 at 0 min, 68/32 at 1.5 min, 3/97 at 21 min, 3/97 at 25 min, 68/32 at 25.1 min with a flow rate of 0.250 mL/min and a temperature of 40°C. The column was re-equilibrated for 2.5 minutes. Detection was carried out in positive and negative ionization modes. MS^2 spectra were collected with fragmentation energies of 10, 20, and 40 eV using MRM-like workflows. LC-MS runs were converted to .mzML files using MsConvert and further processed using the R package `RMassBank` [418]. For each setup, MassBank records were generated independently.

3.5.1.3. Detailed fragmentation studies using DI-MS

Additionally, for LC-MS analysis, lipid reference standards were infused manually using a syringe pump. A working solution of lipids has been prepared by dilution with iPrOH/$CHCl_3$/MeOH + 7.5 mM ammonium formate. Fragmentation studies were performed on the same two MS platforms as above. Collision energies from 10 to 50 eV with 2.5 eV steps were collected for either 1 or 2 minutes. The syringe pump was operated at 500 µL/h or 1000 µL/h in positive and negative modes, respectively. Data was exported to .mzML files using MsConvert, and further processing was performed using the R package `Spectra` [369].

3.5.1.4. Computational resources

Spectra inspection and generation were carried out in R using RStudio and R 4.2. All functions written for parsing and working with lipid shorthand notations have been bundled into a package called `lipidomicsUtils`. This package is available from GitHub (https://github.com/michaelwitting/lipidomicsUtils). Functions for the generation of *in silico* spectra are collected in a package called `wormLipidBlastR`. This package is freely available from GitHub (https://github.com/michaelwitting/wormLipidBlastR). The full functionality is described in the following paragraphs.

3.5.1.5. C. elegans sphingolipid extraction and measurement

C. elegans were grown in liquid culture to obtain sufficient biomass. Worms were grown in S-medium at 20 °C and fed with concentrated *E. coli* NA22. Worms were regularly checked, and *E. coli* NA22 was added to prevent starvation. After one week, worms were harvested and separated from bacteria by filtration using a 2.7 µm Millipore glass fiber filter (Sigma-Aldrich, Taufkirchen, Germany). After two times washing with 10 mL cold M9, worms were frozen at -80 °C until extraction and analysis.

Lipids were extracted according to the Folch method [419]. A pellet of about 750 mg (wet weight) of mixed stage worm samples was mixed with 1 mL of MeOH and homogenized in a Precellys Evolution bead beater (Bertin Instruments, Montigny-le-Bretonneux, France) at about 0 °C and 8000 rpm for three times 10 s with 20 s pause between. After the addition of 2 mL $CHCl_3$, the sample was shaken for one hour at room temperature and 500 rpm using an Eppendorf Thermo- Mixer C (Eppendorf, Hamburg, Germany). Phase separation was induced by the addition of 1 mL H2O and centrifugation at an RCF of 15294 × g and four °C for 15 min. The polar phase was re-extracted with 2 mL of $CHCl_3$ / MeOH / H_2O (v/v, 86 / 14 / 1) for 15 min. After phase separation, organic phases were combined and dried in two aliquots in a SpeedVac Savant centrifugal evaporator (Thermo Scientific, Dreieich, Germany). One aliquot was re-dissolved in 50 µL iPrOH / ACN / H_2O (v/v, 60 / 35 / 5) prior to analysis with UPLC-UHR-ToF-MS and the other in $CHCl_3$ for lipid fractionation.

Sphingolipids were fractionated, according to Bodennec *et al.* [420]. Briefly, different sphingolipids were eluted from either separated or piggy-backed LC-NH_2 (Supelclean LC-NH_2 SPE tubes, 1 mL, 100 mg sorbent, Supelco) and LC-WCX (Supelclean LC-WCX SPE tubes, 1 mL, 100 mg sorbent) SPE columns. Eight different solvent mixtures were used to obtain seven fractions. Fractionation was exactly performed as described in Bodennec et al. [420].

Depletion of glycerol- and glycerophospholipids was adapted from [421]. A 50 µL aliquot of each lipidome fraction was dried and resuspended in 450 µL MeOH and 50 µL 1M KOH in MeOH. The samples were incubated at 37 °C for two h in a Thermomixer and shaken at 1000 rpm. After this time, samples were neutralized with glacial acetic acid (~ 1 µL), 1000 µL $CHCl_3$ was added, and the sample was shaken at 1000 rpm and 20 °C for one h. After the addition of 500 µL H_2O, samples were vortexed, and phases were separated by centrifugation. The lower organic phase was dried in a SpeedVac and redissolved in 50 µL iPrOH / ACN / H_2O (v/v, 60 / 35 / 5) for UPLC-UHR-ToF-MS analysis. Analysis of lipids was performed as described above.

3.5.2. Results and Discussion

3.5.2.1. Analysis of lipid standards

High-quality MS^2 data was generated based on the direct infusion of lipid reference standards and targeted fragmentation using collision energies from 10 to 50 eV in 2.5 eV steps. Spectra were collected for 1 or 2 minutes and exported to .mzML files using MsConvert. All files were then further processed in R using the `Spectra` package. Multiple collected spectra from one precursor and collision energy were combined, and intensities were summed. Peaks needed to be presented at a minimum of 70% of all collected spectra. The resulting spectra were exported to MassBank records. These spectra served as a reference for the construction of the database and definition of fragmentation rules for the covered lipid classes.

The collection of spectra with the stepping collision energy allows identifying optimal collision energies for different tasks. While for non-targeted analysis, a higher number of informative peaks for the identification of lipid species is required, for targeted analysis, the highest possible intensities for quantifier and qualifier ions should be achieved. Plotting fragment intensities against the collision energy can help find optimal collision energies for both tasks. Below, PC 18:0/18:1(9Z) is shown as an example. First, the $[M+H]^+$ was examined. PCs are identified based on the phosphocholine fragment m/z 184.07. In positive ionization mode, this is the main fragment. Sometimes lysophosphatidyl choline fragments are observed. For higher collision energies, additional fragments at *m/z* 104.1 and 125.01 can be observed. These are further fragments of the PC headgroup but don't give additional structural information. Neutral losses of acyl groups are only low in positive ionization mode and can only be detected for highly abundant species.

In negative ionization mode, more structural information can be obtained, e.g. fragments corresponding to the acyl groups bound. However, no position specificity can be achieved. Though the loss of the sn2- position over the sn1-position is preferred, the actual preferences depend on the bound fatty acid and its length and degree of saturation.

3.5.2.2. Parsing of shorthand notations

In order to correctly calculate *m/z* values of lipid fragments for the prediction of spectra, the respective different building blocks need to be determined. Therefore, a package has been created in R called `lipidomicsUtils`. This package can parse a given lipid shorthand notation and determine different groups within the lipid. A similar functionality is available from the R package `rgoslin` [293].

The functionalities are based on the shorthand notation suggested by Liebisch et al. [207, 247]. Besides, differences in carbon length and degree of unsaturation of acyl groups can also show different functional groups. Currently supported additional functional groups are hydroxy groups (OH), keto groups (O), hydroperoxy groups (OOH), and amino groups (NH_2). Based on the shorthand notation the lipid category, main class, and subclass can be

determined. Membership of lipid main classes to different lipid categories is hard coded. Currently, supported lipid categories are fatty acyls (FA, LMFA), glycerolipids (GL, LMGL), glycerophospholipids (GP, LMGP), and sphingolipids (SP, LMSP). Following main lipid classes are mapped to the different categories: FA, CoA, NAE, and PNAE are members of fatty acyls. MG, DG, and TG belong to GL. PC, PE, PS, PG, PGP, PI, PIP, PIP_2, PIP_3, PA, PPA, CL, CDP-DG, and NAPE are members of GP. SPH, Cer, CerP, SM, GlcCer, GalCer, and LacCer are mapped to SP. From the subclasses currently, di- and triacyl, as well as 1-alkyl and 1-alkenyl species, are supported. Four different parsing functionalities for shorthand notations are available: Lipid category, Acyl, alkyl and alkenyl groups, sphingoid bases, and functional groups.

The first block of functions determines the lipid category, main class, or subclass. For example, PC 16:0/18:1(9Z) and PC O-16:0/18:1(9Z) would be classified as GP (glycerophospholipids) as category and as PC in the main class but would yield PC and PC-O as a subclass. This is necessary since fragmentation can differ between these classes. In the next step, building blocks can be isolated from the shorthand notation. Acyl, alkyl, and alkenyl groups in glycero- and glycerophospholipids can be isolated using the following regular expression in R:

```
(m|d|t|O-|P-
)*\\d+:\\d+(\\((\\d*(E|Z|Me|OH|OOH|O|NH2)(\\[(S|R)\\])*,*)*\\))*
```

Based on this, all sphingoid base and acyl groups are returned, and the list is sub-filtered using to not contain `m`, `d`, or `t` (notation of hydroxy group on sphingoid base). This functionality returns all strings that match this regular expression. PC 16:0/18:1(9Z) would return 16:0 and 18:1(9Z) and PC O-16:0/18:1(9Z) return O-16:0 and 18:1(9Z). Likewise, sphingoid bases are isolated with the same regular expression and further filtered to start with either `m`, `d`, or `t`. Further functions allow determining the position and number of different functional groups. Supported functional groups are also isolated as a regular expression. Each functional group has to be accompanied by a number indicating the

position in the respective lipid or acyl, alkyl, or alkenyl, e.g. OH would not be recovered but needs to be defined as 2OH. If the respective group introduces a chiral center to the molecule, this can be defined by adding either R or S in square brackets after the functional group (e.g. 2OH[S]). The functions take any input, but it is recommended to first parse the individual lipids into acyl, alkyl, alkenyl, or sphingoid groups and then perform the detection of functional groups on them individually. Different functions allow the return of the number of a specific functional group or return the complete block (e.g. 16:0(2OH[R]) returns 2OH[R]).

Parsing lipids into their functional groups allows for calculating the mass of individual building blocks and intact lipids. Based on the number of carbons, double bonds, and potentially present functional groups, the number of each element in the acyl, alkyl, and alkenyl groups can be determined. Based on this, a molecular formula is generated from which in turn, the exact mass is calculated. Furthermore, masses of fixed building blocks like the sn-glycerol-3-phosphocholine are hardcoded within the package. This allows the calculation of lipid masses directly from the shorthand notation without the need to generate a molecular structure first.

Furthermore, shorthand notations from lipids at the isomeric subspecies level can be normalized to a common nomenclature. From the shorthand notation of the isomeric subspecies, notations for the structural and the molecular level can be generated using different functions for either fatty acyls, glycerolipids, glycerophospholipids, or sphingolipids. In the case of glycosphingolipids, headgroups are simplified to Hex or Hex2 for mono- or di-hexosyl sugars. The simplification to lower levels of the shorthand notation hierarchy currently only allows the use of unmodified acyl, alkyl, or alkenyl groups, since the rules for handling functional groups like hydroxy groups or others are underdeveloped. For example, the two acyl groups 16:0(2O) and 16:1(2OH[R]) would yield the same mass and cannot be differentiated. However, rules to encode this ambiguity are not yet systematically defined. Different systems have been proposed. One of the best-developed but currently not widely used systems is generated by the epiLION software [422]. Once the rules for handling

functional groups in complex lipids have been standardized by the Lipidomics Standard Initiative, they can be encoded into the `lipidomicsUtils` package.

Lastly, functionalities to calculate adduct masses, ion formulas, and parsing of chemical formulas have been added. Based on the supplied functions, the package is flexible and can handle lipid shorthand notations within the R environment directly without changing to different programming languages.

3.5.2.3. Template spectra

Similar to LipidBlast, the developed spectral prediction tools requires measured template spectra that are used for the generation of *in silico* spectra. These templates can come from published or measured spectra but need to be coded in a specific way to be useful within R. Templates are either stored in text files and can be read or coded into R. They contain instructions on how to calculate the *m/z* value of a fragment and its intensity, scaled to the maximum peak having an intensity of 999. For the here-developed tools, templates are represented as named lists in R, where the names serve as calculation formulas for the *m/z* values of the fragment, and the values represent the respective intensities. The developed tool supplies different fixed building blocks that can be used to calculate fragment *m/z* values, e.g. glycerol backbone, phosphate group, etc. The names of the supplied template list contain the calculation formula and are evaluated in R using the function `eval(parse(text = x ))`. `Adduct_mass` represents the calculated ion mass from the exact mass of the lipid, which is calculated from its shorthand and the supplied adduct. `sn1_mass` and `sn2_mass` (and optionally `sn3_mass` for triacylglycerols) are the masses of the intact, non-charged acyl residues as a fatty acids. Furthermore, different lipid class-specific masses are supplied. For example, in the case of PCs, these include `gpc_mass`, `pc_mass`, and `choline_mass`. Other constant building blocks are `water_mass` or `proton_mass`, for example. The masses of these building blocks are supplied by the lipidomicsUtils package and are hardcoded, and a complete list of all hardcoded building blocks can be found in the appendix (SI Table 7). If a certain constant mass is not supplied, it

can be either invoked by a number or via calculation from a formula using `rcdk` [388]. Since the package is loaded in the background, functions like `get.formula()` can be used to calculate masses from formulas. Variable masses, such as acyl side chains or sphingoid bases, are calculated from the parsed building blocks. The example below shows the template for the fragmentation of PCs as $[M+FA-H]^-$ adducts as it would be constructed.

```
template <- list(
  "adduct_mass" = 10,
  "adduct_mass - rcdk::get.formula('C2H4O2')@mass" = 750,
  "adduct_mass - sn1_mass + water_mass -
rcdk::get.formula('C2H4O2')@mass" = 10,
  "adduct_mass - sn2_mass + water_mass -
rcdk::get.formula('C2H4O2')@mass" = 10,
  "adduct_mass - sn1_mass - rcdk::get.formula('C2H4O2')@mass" = 10,
  "adduct_mass - sn2_mass - rcdk::get.formula('C2H4O2')@mass" = 10,
  "sn1_mass - proton_mass" = 999,
  "sn2_mass - proton_mass" = 999
)
```

Templates can be either directly generated in R or read from text files. The following format for text files was defined. First, the calculation formula similar to the hardcoding in R is supplied, and the intensity is separated from the formula by a ->. If mass calculations based on `rcdk` functions are required, they can also be included. The example above for a PC as $[M+FA-H]^-$ adduct would like in a text file like this:

```
adduct_mass->10
adduct_mass - rcdk::get.formula('C2H4O2')@mass->870
adduct_mass - sn1_mass + water_mass -
rcdk::get.formula('C2H4O2')@mass->150
adduct_mass - sn2_mass + water_mass -
rcdk::get.formula('C2H4O2')@mass->150
adduct_mass - sn1_mass - rcdk::get.formula('C2H4O2')@mass->10
adduct_mass - sn2_mass - rcdk::get.formula('C2H4O2')@mass->10
sn1_mass - proton_mass->999
sn2_mass - proton_mass->999
```

The function `read_template()` reads a template stored in a text file and returns a named list as required by wormLipidBlastR. Different templates are hardcoded and serve as fallback and default templates, if users don't supply their custom templates. They are based on

measurements of lipid standards on a Bruker maXis UHR-ToF-MS at a collision energy of 40eV.

3.5.2.4. Spectra generation

For the generation of spectra for different species, the templates are used together with a shorthand notation on the molecular, structural, or isomer subspecies level. However, since so far differences between sn1 and sn2 are not incorporated into the model, the different levels would yield the same results. Therefore, it is recommended to use the molecular subspecies level (e.g. PC 16:0_18:1).

Based on the supplied shorthand notation, adduct definition, and the respective template, the masses of the different building blocks are calculated and supplied to the function to correct fragment *m/z* values based on the supplied template formula, and a spectrum is generated. With this, a vector of *m/z* values together with a vector of their respective intensities is generated. Spectral data is stored in `Spectra` object from the Spectra R package (https://github.com/rformassspectrometry/Spectra), which can be exported into different formats, such as .mgf, .msp, or MassBank records. All related metadata is stored in defined metadata columns, which include the chemical formula of the molecule, internal ID, or the SPLASH, which is a unique identifier for spectra generated based on the `splashR` package [423]. The currently implemented metadata columns allow the export to a correctly annotated MassBank record. The `export()` function from the MsBackendMassbank allows exporting generated records into valid Massbank record files. The Spectra objects can be concatenated to generate a list of several spectra. Therefore, large libraries can be generated and stored in a single variable which is available for further use in R or can be exported to specific file formats. This allows the direct use of the generated spectra within R, e.g. for comparison against measured spectra using `compareSpectra()` from the Spectra package. Most of the metadata is generated by the code or comes from the user input, e.g. adduct and formula. Users can additionally set the collision energy in the Spectra object if

they want to have a different collision energy set. Additional fields such as RT or CCS can be set optionally.

As benchmarking, predicted lipid species for *C. elegans* were used. In total, spectra for 1545 lipids from different classes were generated within 3322.53 seconds. This means on average, prediction takes about 2 seconds per lipid. However, the code is not parallelized at the current stage, which would further improve speed. Spectra can be generated during each new session based on a list of shorthand notations. However, for reproducible research, it is advantageous to store generated spectra

3.5.2.5. Validation of predicted spectra

Validation of predicted spectra is performed by comparison against well-analyzed and annotated spectra from *C. elegans*. Particularly sphingolipids have been recently investigated in *C. elegans,* and several new species from different sphingolipid classes have been described [116]. The worm uses an unusual C17iso sphingoid base in all its sphingolipids [251, 279]. Since also no reference spectra based on authentic standards exist for these sphingolipids and no structures are present in LipidMaps, which makes the use of LipidFrag not possible, they represent an ideal use case for validation.

Since the exact number of sphingolipids and their structure is unknown, potential lipid structures have been generated *in silico* using the ChemAxon JChem for Excel based on different sphingoid bases and potential fatty acids. Analysis of sphingolipids so far has shown that they typically contain saturated straight chain N-acyl bound fatty acids with or without hydroxylation of the second carbon. A list of suitable fatty acids has been compiled, and in total, 510 sphingolipids at the isomeric subspecies, 324 at the structural or molecular subspecies, and 282 at the species level have been generated. Sphingolipidome data from Hänel *et al.* was used as a validation set [116]. In this data, fragmentation data from 10 and 40 eV have been combined. Fragmentation of mammalian sphingolipids with C18 sphingoid bases was obtained from chemical standards and served as input for the templates (data not shown). Additionally, literature data have been uses, since the fragmentation of sphingolipids

has been extensively studied in the negative ionization mode and only partially in the positive ionization mode [340, 413, 414, 424, 425]. For all predicted sphingolipids, the shorthand notation was generated, and spectra were predicted.

Predicted spectra with a precursor matching within 0.005 Da were chosen for spectral comparison. Data from fractionated C*. elegans* sphingolipids were used for testing. Using these spectra, it was found that dihydro ceramides were not matching well with cosine scores around 0.5, while other sphingolipid classes showed good agreement. This is mostly attributed to differences in the intensities of different fragments. Dihydroceramides are missing the double bound position 4 of the sphingoid base. This double bond adds a mesomery-based stabilizing effect. Therefore, the intensity of fragments is in DhCer lower compared to ceramides. Since no dihydro ceramides were included in the templates, this mismatch can be explained. Spectra of some specific matches have been examined in more detail. Figure 19 shows an example of a ceramide fragmentation spectrum obtained from *C. elegans* samples. This spectrum is a composite spectrum of two collision energies (10 eV and 40eV) generated and merged during acquisition with DDA. The highest fragment peak is the loss of a water molecule from the precursor. This fragmentation usually occurs at low collision energies. Typical fragments for Cers are the loss of a water molecule, which is possible for both hydroxyl groups. This fragmentation requires low energy, since it is often already observed as an in-source fragment. Both obtained structures can lose the N-Acyl, again yielding fragments of similar masses. One of these fragments is losing an additional water molecule, while the other is losing formaldehyde. The fragments yield a typical peak triplet with the *m/z* 238.2529, 250.2529 and 268.2635. These three fragments can identify a molecule as a sphingolipid containing a C17:1 sphingoid base. The N-Acyl can be identified by the difference between the $[M\text{-}H_2O+H]^+$ fragment and *m/z* 268.2635 (see Figure 19). In contrast to Cers, DhCers show the typical water loss to a lesser extent. This suggests that the loss of the hydroxyl group at position 3 is the favored one in ceramides because of the stabilizing effect of the neighboring double bond. Additionally, *m/z* 288.2897 is observed for DhCers, which represents the loss of the fatty acyl as ketene directly from the $[M+H]^+$. In

DhCers, this double bond is missing; therefore DhCers tend to show lower intensities for the $[M-H_2O+H]^+$ fragment. Interestingly, DhCers show an N-Acyl fragment corresponding to an N-Acyl amide. These fragments can also be found in highly abundant Cer fragmentation spectra, but only at low intensities. Similar N-Acyl amide fragments as the ones observed for DhCers were found in GlcCers. GlcCers showed different fragments related to the loss of the hexosyl headgroup. Neutral losses of 162.0528 Da were observed from the $[M+H]^+$ and $[M-H_2O+H]^+$ peaks. DhSMs and SMs yield a fragment of *m/z* 184.0733 corresponding to the phosphocholine headgroup. This was the only observed fragment, and therefore no further structural characterization was possible for SMs. Examples from matching results are summarized in Table 24.

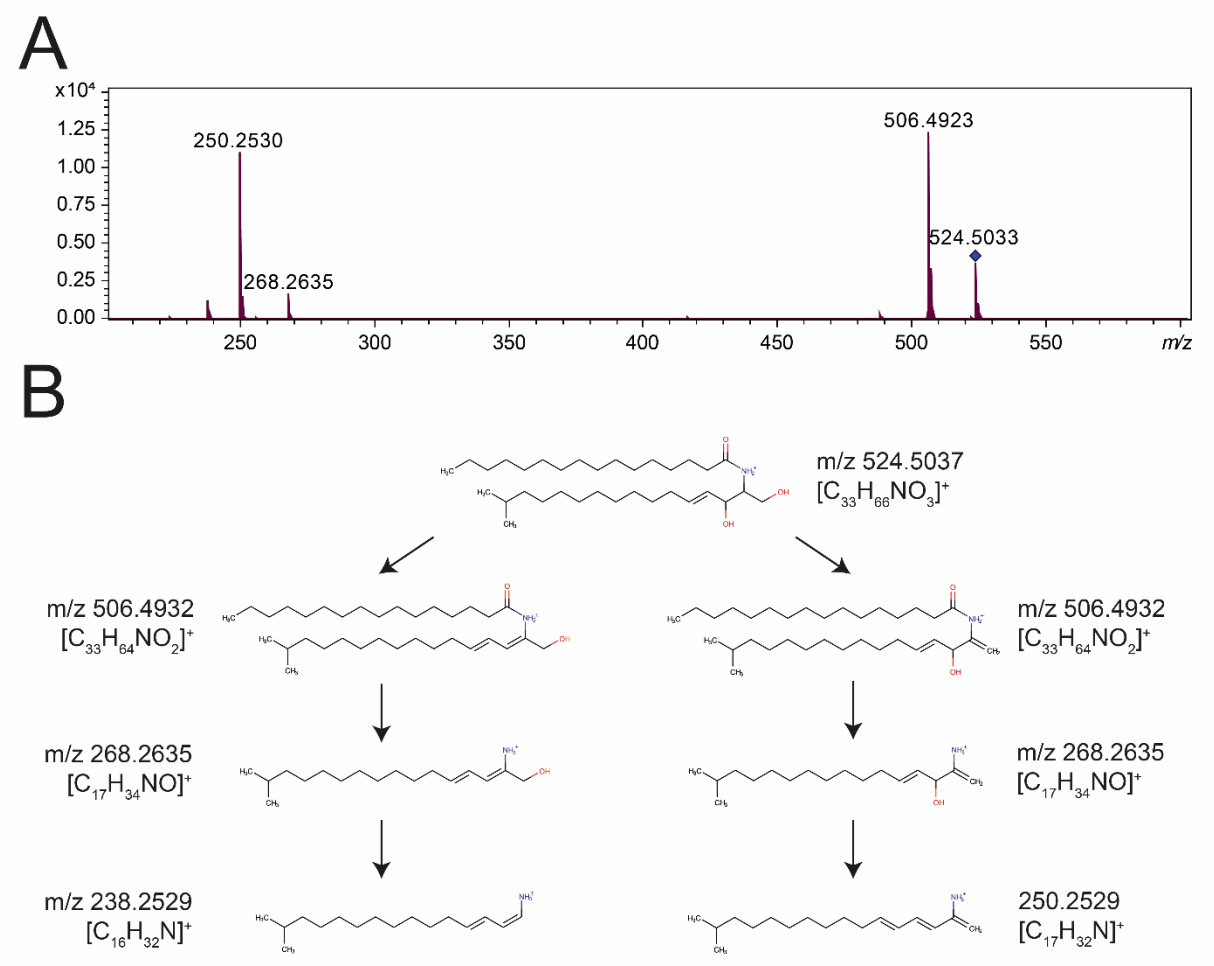


Figure 19: (A) Example spectrum for a C. elegans ceramide (Cer d17:1/16:0) (B) Fragmentation pathway of Cer d17:1/16:0

Table 24: Results from spectral matching using confirmed sphingolipid identifications from Hänel et al. [116] as query spectra and spectra predicted by wormLipidBlastR based on the lipids predicted in chapter 2.3 as target database. Differences in nomenclature for the lipids are based on the reporting. Full structural details are reported for the WormLipidBlastR results since they were used for spectra generation.

Sphingolipid from Hänel *et al.* [116]	Best Matching lipid from wormLipidBlastR	Forward Score	Revere Score
Cer d17:0/22:0	Cer 16:0(3OH,4OH,15Me)/22:0	488	524
Cer d17:0/22:0(OH)	Cer 16:0(3OH,4OH,15Me)/22:0(2OH[R])	536	568
Cer d17:0/23:0	Cer 16:0(3OH,4OH,15Me)/23:0	514	545
Cer d17:0/23:0	Cer 16:0(3OH,4OH,15Me)/22:0(21Me)	514	545
Cer d17:0/24:0 (isomer 1)	Cer 16:0(3OH,4OH,15Me)/24:0	556	619
Cer(d17:0/24:0 (isomer 2)	Cer 16:0(3OH,4OH,15Me)/24:0	533	577
Cer d17:0/26:0	Cer 16:0(3OH,4OH,15Me)/26:0	492	537
Cer d17:1/16:0	Cer 16:1(4E,3OH,4OH,15Me)/16:0	969	993
Cer d17:1/18:0	Cer 16:1(4E,3OH,4OH,15Me)/18:0	968	994
Cer d17:1/20:0	Cer 16:1(4E,3OH,4OH,15Me)/20:0	972	994
Cer d17:1/20:0(OH) (isomer 1)	Cer 16:1(4E,3OH,4OH,15Me)/20:0(2OH[R])	932	982
Cer d17:1/22:0(OH) (isomer 2)	Cer 16:1(4E,3OH,4OH,15Me)/22:0(2OH[R])	955	976
Cer d17:1/22:0(OH)	Cer 16:1(4E,3OH,4OH,15Me)/22:0(2OH[R])	952	995
Cer d17:1/24:0 (isomer 1)	Cer 16:1(4E,3OH,4OH,15Me)/24:0	933	968
Cer d17:1/24:0 (isomer 2)	Cer 16:1(4E,3OH,4OH,15Me)/24:0	949	974
Cer d17:1/24:0(OH) (isomer 1)	Cer 16:1(4E,3OH,4OH,15Me)/24:0(2OH[R])	949	988
Cer d17:1/24:0(OH) (isomer 2)	Cer 16:1(4E,3OH,4OH,15Me)/24:0(2OH[R])	878	965
Cer d17:1/26:0	Cer 16:1(4E,3OH,4OH,15Me)/26:0	921	945
Cer d17:1/26:0(OH) (isomer 1)	Cer 16:1(4E,3OH,4OH,15Me)/26:0(2OH[R])	959	994
Cer d17:1/26:0(OH) (isomer 2)	Cer 16:1(4E,3OH,4OH,15Me)/26:0(2OH[R])	937	987

Sphingolipid from Hänel *et al.* [116]	Best Matching lipid from wormLipidBlastR	Forward Score	Revere Score
HexCer d17:1/21:0(OH) (isomer 1)	GlcCer 16:1(4E,3OH,4OH,15Me)/21:0(2OH[R])	843	951
HexCer d17:1/21:0(OH) (isomer 2)	GlcCer 16:1(4E,3OH,4OH,15Me)/20:0(2OH[R],19Me)	843	951
HexCer d17:1/23:0(OH) (isomer 1)	GlcCer 16:1(4E,3OH,4OH,15Me)/23:0(2OH[R])	860	972
HexCer d17:1/23:0(OH) (isomer 2)	GlcCer 16:1(4E,3OH,4OH,15Me)/22:0(2OH[R],21Me)	860	972
HexCer d17:1/24:0(OH) (isomer 1)	GlcCer 16:1(4E,3OH,4OH,15Me)/24:0(2OH[R])	780	939
HexCer d17:1/24:0(OH) (isomer 2)	GlcCer 16:1(4E,3OH,4OH,15Me)/24:0(2OH[R])	799	947

3.5.3. Conclusion

The framework developed for the prediction of lipid MS^2 spectra for *C. elegans* represents a modular system, which can be easily expanded to further lipid classes. While the original LipidBlast implementation relies on an Excel sheet, which is useful for users with limited programming experience, the customization of spectra required some serious effort in reprogramming parts of the underlying VBA code. The presented implementation in R is easy to handle and allows customization of predicted lipid spectra based on a building block approach. Furthermore, new libraries can be created during data analysis using the R code within a running session, which offers higher flexibility. Users can define their own calculation rules for the calculation of fragment m/z values. Further lipid classes can be added at any time. A prerequisite for successful integration is that the package `lipidomicsUtils` can parse the new lipid class. Therefore, a defined shorthand notation as well as rules to derive building blocks from them, are required. The further development of the framework and `lipidomicsUtils` go hand in hand.

A potential additional feature for future integration is the inclusion of intensity modeling to predict the intensity of acyl-derived fragments correctly. For example, Schumann *et al.*

recently provided models for data obtained on Orbitrap instrumentation to harmonize the abundances of fatty acid anion fragments derived from the sn1 and sn2 position of glycerophospholipids [426]. Based on their model, it was possible to adjust after the data acquisition for instruments settings and collision energy. Since `lipidomicsUtils` is able to supply further information of building blocks like the number of double bonds or their position, similar modeling can be achieved. However, several lipid standards must be measured systematically in order to obtain valid models.

Another potential application for further improvement is the prediction of spectra for oxidized lipids. Specialized software for the prediction of spectra of these lipids exists, but no generic framework that integrates non-oxidized and oxidized lipids in one package exists [427].

3.6. Integrated annotation and identification pipeline

Using different available information, such as *m/z*, RT, CCS, and fragmentation data, metabolites and lipids can be annotated and identified on different levels. Annotation of measured *m/z* values is still one of the first steps in metabolite annotation for metabolites not covered in in-house databases. First, an exact mass search can be performed by comparing measured *m/z* values are compared against theoretical *m/z* values calculated from a compound database and adduct definitions. The MassTRIX server was one of the first tools for exact mass searching and was initially developed for the annotation of ultra-high resolution MS data coming from FT-ICR-MS-based measurements [194]. While the next version also allowed the annotation of LC-MS-based data and integration of transcriptomics, it was still only available as a web server [195]. Another tool is Paintomics, which offers functionalities similar to MassTRIX [197, 428]. Likewise, the tool CEU Mass Mediator follows the same line, but the latest version also integrates MS^2 spectra matching [429, 430].

However, annotation solely based on *m/z* values is very weak since several potential isomeric and isobaric structures for a single *m/z* might exist. Several tools have been developed to overcome this issue. The tool metDNA integrates annotation with a metabolic reaction network using well-identified metabolites as seeds for the annotation of further metabolites [431]. Multiple recursive rounds are performed, and each improves annotation. Another application for the exploration of MS^2 data is metCirc [432]. It groups metabolites based on the similarity of MS^2 spectra and compound familial groupings.

Once the number of potential candidates has been narrowed down, tandem MS is applied to identify metabolites of interest. Dereplication of previously identified metabolites is a critical issue to not re-identify the same metabolites repeatedly. Typically, reference databases are constructed using metabolite standards, and their MS^1 and MS^2 spectra and retention behavior, are stored in an in-house database. Metabolites from the database can be then identified by matching *m/z* values, RTs, and fragmentation spectra of measured metabolites against the reference database.

In order to enable large-scale automatic annotation on the MS^1 and MS^2 levels an R-based workflow has been implemented. This workflow is based on functions implemented in different packages from the RforMassSpectrometry environment. Most of these functions have been implemented in `MetaboCoreUtils` and `MetaboAnnotation` [369, 433, 434].

3.6.1. Material and Methods

3.6.1.1. Implementation

Different functionalities for the annotation rely on existing packages from the RforMassSpectrometry environment, specifically on the `Spectra`, `MetaboCoreUtils`, and `MetaboAnnotation` packages as well as `MsBackendMassBank`, `MsBackendMgf`, and `MsBackendMsp` for the import of measured and library spectra [369]. Inputs are provided in the form of a .yaml file containing paths to all input data and folder and all necessary parameters to perform the annotation, e.g. ppm or absolute errors, RT tolerances and minimum dot-product, etc. The workflow is freely available from GitHub (https://github.com/michaelwitting/MetaboliteAnnotationWorkflow), contains all required functions, and performs installation of all required packages. Additionally, test data is available for testing the workflow without the need to provide own data. The complete workflow is described in the result section.

3.6.2. Results and Discussion

The metabolite annotation server MassTRIX previously worked with relatively static databases that required an administrator to update them if changes occurred, which lowers its flexibility. Furthermore, filtering to specific classes of metabolites has been done post-annotation. This is often time-consuming since the full annotation must be performed all the time. Flexibility was increased by allowing the input of an arbitrary data frame containing metabolites of interest and their specific *m/z* values. Additionally, since annotation on the MS^1 level results in very ambiguous results, MS^2 spectra matching has been integrated into the new workflow. The developed workflow relies on several blocks of code performing different tasks. The individual blocks are discussed below. The presented workflow is based

on functions available from different packages from the RforMassSpectrometry consortium. Since the defined input consists of structured folders and data, it can also be used in workflow tools.

3.6.2.1. MS^1 and MS^2 data import

In the first step, data is imported from different sources. MS^1 data can be imported in the form of simple .tsv or .csv files. Positive and negative ionization mode data are stored in separate QFeatures objects within R. Since data often contains isotope cluster and adduct group information, but no MS^1 spectra are exported from different software tools such as Genedata Expressionist for MS, a function that reconstructs the MS^1 isotope pattern from ungrouped and grouped has been added. MS^1 spectra are returned as `Spectra` objects with one averaged isotope pattern per spectrum. In case MS^1 spectra are available, they can be imported from .mgf files, and the reconstruction step is skipped.

MS^2 data is imported from .mgf or .msp files. In order to be able to link MS^1 and MS2 data, either unique identifiers for the MS^1 feature are created, or existing identifiers are used to flag the corresponding MS^2 spectra with them. This happens by either using existing correspondence between the level stored in the files or by checking if precursor *m/z* and RT values fall into a specific region around an MS^1 feature *m/z* and RT.

3.6.2.2. MS^1 metabolite annotation

One of the core functionalities of MassTRIX or CEU Mass Mediator is to annotate *m/z* values with potential metabolites using databases of theoretical *m/z* values from different metabolites. This function is further available in the newly developed workflow, though it has to be mentioned that this way of annotation yields highly ambiguous annotations with many false positives. Annotation on the MS^1 level is based on the `matchMz()` function from the `MetaboAnnotation` package. As Input is the `rowData` from the respective `QFeatures` objects, as well as databases with metabolites, is used. Annotation using multiple compound databases is possible since several databases can be stored in the input folder. Parameters for the annotation, such as *m/z* or RT tolerance, are read from the .yaml file storing all

configuration parameters. Two cases can be distinguished, either in-house databases, which contain RTs, or external compound databases, with no RT information. Data is supplied as a simple .tsv file with the columns `id`, `name`, `formula`, `exact_mass`, and optionally `rt`. Only neutral compounds shall be provided since the calculation of *m/z* is performed during the annotation process. Furthermore, valid adduct names according to the adduct names defined in `MetaboCoreUtils` are required. Based on the type of library (in-house or external), a parameter object for the `matchMz` function is constructed. If an in-house library is used, a `Mass2MzRtParam` is created; otherwise, a `Mass2MzParam`. Both objects give instructions to the `matchMz` function on how the annotation shall be performed. In both cases, neutral masses are converted to *m/z* values using the adduct definitions. In contrast to previously used static *m/z* tables, this allows more flexibility in the annotation process. Results are stored in dedicated output folders for external or in-house libraries. Data is stored as .tsv or .rds files for further use in the downstream data analysis workflows. Table 25 summarizes all currently supported adducts for annotation.

Table 25: Summary of all currently supported adducts in `MetaboCoreUtils`. The adduct list has been adopted from Huang et al. [435]. Adducts are stored a list that is read upon loading of the package and can be extended anytime. Calculation is split into a multiplicative and an additive part.

Ionization mode	Common Adduct Name	`MetaboCoreUtils` adduct name	Charge	Calculation
positive	$[M+3H]^{3+}$	[M+3H]3+	3+	M / 3 + 3 * 1.007276 / 3
	$[M+2\ H+Na]^{3+}$	[M+2H+Na]3+	3+	M / 3 + (2 * 1.007276 + 1 * 22.98922) / 3
	$[M+H+2\ Na]^{3+}$	[M+H+2Na]3+	3+	M / 3 + (1 * 1.007276 + 2 * 22.98922) / 3
	$[M+3Na]^{3+}$	[M+3Na]3+	3+	M / 3 + (3 * 22.98922) / 3
	$[M+2\ H]^{2+}$	[M+2H]2+	2+	M / 2 + (2 * 1.007276) / 2
	$[M+H+NH4]^{2+}$	[M+H+NH4]2+	2+	M / 2 + (1.007276 + 18.03383) / 2
	$[M+H+K]^{2+}$	[M+H+K]2+	2+	M / 2 + (1.007276 + 38.96316) / 2
	$[M+H+Na]^{2+}$	[M+H+Na]2+	2+	M / 2 + (1.007276 + 22.98922) / 2
	$[M+ACN+2H]^{2+}$	[M+C2H3N+2H]2+	2+	M / 2 + (2 * 1.007276 + 41.02655) / 2

Ionization mode	Common Adduct Name	MetaboCoreUtils adduct name	Charge	Calculation
	[M+2 Na]$^{2+}$	[M+2Na]2+	2+	M / 2 + 2 * 22.98922 / 2
	[M+2 ACN+2 H]$^{2+}$	[M+C4H6N2+2H]2+	2+	M / 2 + (2 * 1.007276 + 2 * 41.02655) / 2
	[M+3 ACN+2 H]$^{2+}$	[M+C6H9N3+2H]2+	2+	M / 2 + (2 * 1.007276 + 3 * 41.02655) / 2
	[M+H]$^{+}$	[M+H]+	1+	M + 1.007276
	[M+Li]$^{+}$	[M+Li]+	1+	M + 7.015456
	[M+2 Li-H]$^{+}$	[M+2Li-H]+	1+	M + 2 * 7.015456 - 1.007276
	[M+NH4]$^{+}$	[M+NH4]+	1+	M + 18.03383
	[M+H_2O+H]$^{+}$	[M+H2O+H]+	1+	M + 19.01784
	[M+Na]$^{+}$	[M+Na]+	1+	M + 22.98922
	[M+CH3OH+H]$^{+}$	[M+CH4O+H]+	1+	M + 1.007276 + 32.02621
	[M+K]$^{+}$	[M+K]+	1+	M + 38.96316
	[M+ACN+H]$^{+}$	[M+C2H3N+H]+	1+	M + 1.007276 + 41.02655
	[M+2 Na-H]$^{+}$	[M+2Na-H]+	1+	M + 2 * 22.98922 - 1.007276
	[M+2K-H]$^{+}$	[M+2K-H]+	1+	M + 2 * 38.96316 - 1.007276
	[M+iPrOH+H]$^{+}$	[M+C3H8O+H]+	1+	M + 1.007276 + 60.05751
	[M+ACN+Na]$^{+}$	[M+C2H3N+Na]+	1+	M + 22.98922 + 41.02655
	[M+DMSO+H]$^{+}$	[M+DMSO+H]+	1+	M + 1.007276 + 78.01394
	[M+2 ACN+H]$^{+}$	[M+C4H6N2+H]+	1+	M + 1.007276 + 2 * 41.02655
	[2M+H]$^{+}$	[2M+H]+	1+	2 * M + 1.007276
	[2M+NH4]$^{+}$	[2M+NH4]+	1+	2 * M + 18.03383
	[2M+Na]$^{+}$	[2M+Na]+	1+	2 * M + 22.98922
	[2M+K]$^{+}$	[2M+K]+	1+	2 * M + 38.96316
	[2M+ACN+H]$^{+}$	[2M+C2H3N+H]+	1+	2 * M + 1.007276 + 41.02655

Ionization mode	Common Adduct Name	`MetaboCoreUtils` adduct name	Charge	Calculation
	$[2M+ACN+Na]^+$	[2M+C2H3N+Na]+	1+	2 * M + 22.98922 + 41.02655
	$[3M+H]^+$	[3M+H]+	1+	3 * M + 1.007276
negative	$[M-3H]^{3-}$	[M-3H]3-	3-	M / 3 – 3 * 1.007276 / 3
	$[M-2H]^{2-}$	[M-2H]2-	2-	M / 2 – 2 * 1.007276 / 2
	$[M-H]^-$	[M-H]-	1-	M – 1.007276
	$[M+Na-2H]^-$	[M+Na-2H]-	1-	M - 2 * 1.007276 + 22.98922
	$[M+Cl]^-$	[M+Cl]-	1-	M + 34.9694
	$[M+K-2H]^-$	[M+K-2H]-	1-	M - 2 * 1.007276 + 38.96316
	$[M+ACN-H]^-$	[M+C2H3N-H]-	1-	M + 41.02655 - 1.007276
	$[M+FA-H]^-$	[M+CHO2]-	1-	M + 44.9982
	$[M+HAc-H]^-$	[M+C2H3O2]-	1-	M + 59.01385
	$[M+Br]^-$	[M+Br]-	1-	M + 78.91889
	$[M+TFA-H]^-$	[M+C2F3O2]-	1-	M + 112.9856
	$[2M-H]^-$	[2M-H]-	1-	2 * M - 1.007276
	$[2M+FA-H]^-$	[2M+CHO2]-	1-	2 * M + 44.9982
	$[2M+HAc-H]^-$	[2M+C2H3O2]-	1-	2 * M + 59.01385
	$[3M-H]^-$	[3M-H]-	1-	3 * M – 1.007276

3.6.2.3. MS^2 data processing, library search, and spectra analysis

MS^2 spectral library matching represents the current gold standard for metabolite annotation. Using in-house libraries, annotation can be performed based on the precursor *m/z*, RT, and spectral similarity resulting in high confidence annotations. For metabolites not present in the in-house database matching against reference spectra from other databases represents an additional alternative for annotation.

Similar to the MS^1 annotation, the MS^2 annotation decides on the origin of the library if additional RT matching shall be performed. A comparison of spectra is performed using the `matchSpectra` function from `MetaboAnnotation`. Spectral comparison is performed using the forward and reverse dot products. Libraries can be supplied as .mgf, .msp, or MassBank records. If the latter is used, all spectra must be supplied in one file with all records. Both query and target spectra are normalized according to their base peak to have a maximum intensity of 100. Additionally, low-intensity peaks, which might interfere with comparison, can be removed using an intensity threshold supplied with all other parameters. Library spectra are filtered using the precursor *m/z* of the query spectrum before the actual spectral comparison is performed. The actual spectral similarity is calculated using the normalized dot-product described by Scott and Stein [436].

$$DP = \frac{(\sum W_{act.} W_{lib})^2}{\sum W_{act}^2 \sum W_{lib}^2}$$

For matching, the forward and reverse dot products are calculated. The forward dot-product takes all peaks into account, while the reverse dot-product only uses peaks matching to library peaks. In addition, the number of matching peaks is reported. Results from each individual spectrum are combined, and a data frame with all results is returned. If matching against an in-house library containing RT information is also performed, the RT difference between query and library spectra is reported.

The results of matching can be stored as .tsv or .rds files. The latter has the advantage that they contain the original `matchSpectra` object, which can be loaded again into R for further processing and data evaluation. This object contains all the query and target spectra used. Manual evaluation of spectral matching results can be performed using the `validateMatchedSpectra` function from `MetaboAnnotation`, which opens as a Shiny app that can display all query spectra and their associated results. This allows manual validation of all results and improves data quality.

3.6.2.4. General utility functions

Besides the annotation of MS1 and MS2 data, several additional functions are available. These include, for example, the matching of features between ionization modes. If the same chromatographic system is used in positive and negative ionization modes and metabolites can be ionized in both ionization modes, RTs can be used to match between them. This would allow the comparison and the transfer of annotations between the ionization modes. The `matchMz` function is also used for this, together with the `Mz2MassParam` or `Mz2MassRtParam` object. Using the parameter objects, *m/z* values are converted to neutral masses for all defined adducts in positive and negative ionization modes. The neutral masses are then compared against each other for all combinations of adducts from both ion modes. Matches are stored with the adduct combination as a string for further processing. Feature-based molecular networking represents a powerful tool alternative to molecular networking [437]. The main improvement is that MS^2 spectra of features with the same precursor *m/z* but different RTs are no longer merged together, and therefore, the complete networking becomes aware of isomeric structures. The developed workflow contains export functions that generate files that can be directly uploaded to GNPS into the FBMN workflow. Furthermore, upon import of MS^1 data, isotope patterns are reconstructed and stored in a separate `Spectra` object. Both MS^1 and MS^2 data are used to export Sirius .ms files, which can be directly imported into Sirius for fragmentation tree calculation and spectral analysis.

3.6.2.5. Utility functions for lipidomics

Besides the annotation workflow, some utility functions specifically for the analysis of lipids have been developed. The two functions `containsMz()` and `containsNeutralLoss()`, allowing to search for certain *m/z* values that are typical for lipids (examples are listed in Table 26). Both functions would return a vector of `TRUE` or `FALSE` if a matching *m/z* value was found in the `Spectra` object. These functions have been implemented in the `Spectra` package. Furthermore, for identifying different lipid classes, functions to work with referenced Kendrick mass defects (RKMDs) have been integrated. This approach has been developed

by Lerno *et al.* and first calculates the Kendrick mass defect and then references this to mass defects of lipid classes [438] (compare chapter 3.3.2.5). The functions `calculateKm()`, `calculateKmd()`, and `calculateRkmd()` perform the respective steps and have been integrated into the `MetaboCoreUtils` package.

Table 26: Lipid class specific product ions and neutral losses that can be used in conjunction with the `containsMz` and `containsNeutralLoss` functions from the `Spectra` package. Masses for sphingolipids assume a C17 sphingoid base.

Ionization mode	Lipid class	Adduct	Fragment / neutral loss
positive	PC	$[M+H]^+$	*m/z* 184.0733 NL 183.0661
		$[M+Na]^+$	*m/z* 146.984 NL 183.0661
	PE	$[M+H]^+$	NL 141.0191
		$[M+Na]^+$	*m/z* 120.9661 *m/z* 150.005 NL 44.0392
	PS	$[M+H]^+$	NL 185.0089
		$[M+Na]^+$	*m/z* 207.998145
	PG	$[M+H]^+$	NL 172.0137
		$[M+Na]^+$	*m/z* 195.0029
		$[M+NH_4]^+$	NL 189.0402
	PA	$[M+H]^+$	NL 97.976896
		$[M+Na]^+$	---
		$[M+NH_4]^+$	NL 115.0035 NL 17.0266
	PI	$[M+H]^+$	NL 260.0297
		$[M+Na]^+$	---
		$[M+NH_4]^+$	NL 277.0563
	DG	$[M+Na]^+$	---
		$[M+NH_4]^+$	NL 35.0371 NL 17.0266

Ionization mode	Lipid class	Adduct	Fragment / neutral loss
	SM	$[M+H]^+$	*m/z* 184.0733
	Cer	$[M+H]^+$	NL 18.0106 *m/z* 238.2529 *m/z* 250.2529 *m/z* 268.2635
	HexCer	$[M+H]^+$	NL 18.0106 NL 162.0528 *m/z* 238.2529 *m/z* 250.2529 *m/z* 268.2635
	Hex2Cer	$[M+H]^+$	NL 18.0106 NL 162.0528 NL 342.1162 *m/z* 238.2529 *m/z* 250.2529 *m/z* 268.2635
Negative	PC	$[M+FA-H]^-$	NL 60.0211
	PE	$[M-H]^-$	*m/z* 140.0118 *m/z* 196.0380
	PG	$[M-H]^-$	*m/z* 152.9958 *m/z* 171.0063
	PA	$[M-H]^-$	*m/z* 152.9958
	PI	$[M-H]^-$	*m/z* 241.0119 *m/z* 223.0013 *m/z* 152.9958 *m/z* 259.0224
	PS	$[M-H]^-$	*m/z* 152.9958 NL 87.0320
	SM	$[M+FA-H]^-$	*m/z* 168.0431 NL 60.0211

3.6.3. Conclusion

The developed annotation workflow allows a first-line automatic annotation of LC-MS/MS data on the MS^1 and MS^2 levels as well as several utility functionalities. It is based on the package from the RforMassSpectrometry environment and can handle positive and negative mode data from the same experimental setup. Export of results to multiple file formats as

well as export data for external tools, allows easy integration. This workflow is used to annotate metabolomics and lipidomics data in the following chapters. The workflow also integrates other external tools, such as GNPS or Sirius, by exporting data in the required format. Since consistent identifiers are generated and used for each feature, export results can be cross-matched between the tools. In the future the integration of additional tools such as MetFrag is possible.

4. Large-scale identification of lipids in *C. elegans*

Chapter-related publications:

HLH-30 dependent rewiring of metabolism during starvation in C. elegans

Dall K. B., J. F. Havelund, E. B. Harvald, M. Witting, N. J. Færgeman

Aging Cell, 2021 Apr;20(4):e13342. doi: 10.1111/acel.13342.

This article describes the role of the transcription factor HLH-30/TFEB in the starvation response of *C. elegans*. Metabolomic and lipidomic analysis has been performed to study starvation-related changes in WT and *hlh-30* mutants. In collaboration with the group from Odense, I performed lipidomics analysis, including lipid annotation and statistical analysis. This is one of three datasets used in this chapter for in-depth lipid annotation.

Reduced peroxisomal import triggers a peroxisomal retrograde signaling

Rackles E., I. Forné, C. Fischer, X. Zhang, S. Schrott, J. Zacherl, M. Witting, J. Ewbank, C. Osman, A. Imhof, S. G. Rolland

Cell Reports, 2021 Jan 19;34(3):108653. doi: 10.1016/j.celrep.2020.108653.

Lipidomics was used to study the effect of knocking down *prx-5*/PEX5, causing peroxisomal import stress. In the lipid analysis of *C. elegans* samples, lipid annotation, and statistical analysis were performed by myself. This is one of three datasets used in this chapter for in-depth lipid annotation.

Autophagy compensates for defects in mitochondrial dynamics

Haeussler S., F. Köhler, M. Witting, M. F. Premm, S. G. Rolland, C. Fischer, L. Chauve, O. Casanueva, B. Conradt

PLoS Genetics, 2020 Mar 19;16(3):e1008638

Mitochondrial dynamics are essential for mitochondrial and cellular homeostasis. Lipidomic analysis revealed changes in triacylglycerols with specific chain length and degree of desaturation. I performed lipid analysis including lipid annotation and statistical analysis. This is one of three datasets used in this chapter for in-depth lipid annotation.

Comparison of lipidome profiles of Caenorhabditis elegans – Results from an inter-laboratory ring trial

Spanier B., A. Laurençon, A. Weiser, N. Pujol, S. Omi, A. Barsch, S. W. Meyer, J. J. Ewbank, F. Paladino, S. Garvis, H. Aguilaniu, M. Witting

Metabolomics, 2021 Feb 17;17(3):25. doi: 10.1007/s11306-021-01775-6.

Differences between metabolomics and/or lipidomics results from different laboratories can be attributed to differences in the culture of *C. elegans*, e.g. different nutritional value of feeding bacteria. The aim of this study was to find out how strong such differences are reflected in lipidomics results. I planned and oversaw the entire research and performed lipid analysis, annotation, and statistical analysis.

4.1. Introduction

Lipids play important roles in different aspects of *C. elegans* biology. However, so far, no *C. elegans*-specific lipidome database exists. The analysis of C. elegans lipid metabolism until now was mostly driven by the analysis of fatty acids based on the methodology and pathway delineated by Watts and Browse [252]. Beside GC-MS, several other methods have been used for the analysis of lipid-related phenotypes or lipid read-outs. Since the worm is transparent, different techniques can quickly assess internal lipid storage. Nile red, a phenoxazone dye or BODIPY-labeled fatty acids, can be used for imaging lipid deposits [439, 440]. However, analysis has shown that these dyes do not label major fat stores in *C. elegans,* and oil red O should be used instead, which shows good agreement with biochemical methods [441]. Several other techniques for imaging live animals exist [442, 443]. Lipidomics has been recently applied to different fields of *C. elegans* biology. MS-based lipid analysis in *C. elegans* is now relatively advanced, and various methods are used to access different lipid classes and obtain results [68, 130]. LC-MS-based lipidomics is the most used analysis method to analyze lipids from the worm [68, 130]. However, also shotgun analysis is used [100, 113, 154, 277]. No systematic investigation to define the *C. elegans* lipidome has been performed, which a few exceptions with a few exceptions. One example are sphingolipids, which have been recently investigated in a systematic manner. Eighty-two different sphingolipids have been detected in mixed-stage *C. elegans* cultures, including 18 that have not been detected before [116]. Another example is the in-depth investigation of maradolipids using either shotgun lipidomics or LC-IMS-MS [80, 154].

Lipidome atlases and collections are becoming increasingly important in defining the lipidomes of different cells or organisms. They serve as a reference for future investigations and advance the current knowledge on lipids and their regulation. One particular example is AdipoAtlas, published by Lange *et al.* [444]. White adipose tissue from several lean and obese people has been analyzed using different analytical approaches to define the lipidome of this tissue. Over 1500 lipid species have been identified and semi-quantified. Symons *et al.* performed the analysis of different mammalian cell membranes to define a lipidomic atlas

[445]. In another example, Oemer *et al*. performed an analysis of cardiolipins between different organisms and tissues [296]. Lastly, the HMDB contains a large number of lipids in the latest version and can also be regarded as a reference for the human lipidome [87].

In order to construct a first blueprint of the *C. elegans* lipidome, lipids either predicted by the workflow developed in chapter 2.3, curated from literature, or detected in different datasets created in the last years were compared. Three datasets from different laboratories and publications have been reprocessed with a workflow for detailed lipid curation and analyzed in detail. Using different annotation tools and databases, as many lipids as possible have been identified and manually verified. Generated reference lists were compared against the literature to identify lipids species stably identified in *C. elegans* forming the core lipidome of the nematode.

4.2. Lipid identification in LC-MS/MS datasets

4.2.1. Material and Methods

4.2.1.1. Chemicals

HPLC-grade methyl-tert-butyl ether (MTBE), HPLC-grade chloroform ($CHCl_3$), LC-MS-grade methanol (MeOH), iso-propanol (iPrOH), acetonitrile (ACN), ammonium formate, and formic acid were obtained from Sigma-Aldrich (Sigma-Aldrich, Taufkirchen, Germany). Water was purified using a Merck Millipore Integral water purification system with a resistance of 18 MΩ and TOC < 5 ppb.

4.2.1.2. C. elegans samples

C. elegans samples from different studies have been used, and the BUME extraction with some minor modifications was used [165, 168, 187, 446]. Briefly, 50 µL ice-cold methanol was added to each sample and transferred to beat-beating tubes (NucleoSpin Bead Tubes Type A, Macherey Nagel). The samples were beat beaten three times 10 s with 20 s pause in a Precellys Beat Beating system (Bertin Technologies). The additional Cryolys module was used with liquid nitrogen to prevent excessive heating of samples during disruption. 150 µl butanol and 200 µl heptane/ethyl acetate (3:1, v/v) were added to each sample sequentially, which were then incubated for 1 h at 500 rpm/RT. 200 µl 1% acetic acid in H2O was added to each sample, followed by centrifugation for 15 min at 17949 × g/4°C. The upper organic phase was transferred to a fresh Eppendorf tube, and the lower aqueous phase was re-extracted by the addition of 200 µl heptane/ethyl acetate (3:1, v/v) followed by incubation and centrifugation as described above. The upper organic phase was combined with the previously obtained organic phase. Samples were evaporated to dryness and stored at −20°C until further analysis. Samples were re-dissolved in 50 µl H_2O/ACN/iPrOH (5/35/60, v/v/v) vortexed, and 40 µl were transferred to an autosampler vial. The remaining 10 µl were pooled to form a QC sample for the entire study. The precipitated proteins were used for the determination of protein content using a Bicinchoninic Acid Protein Assay Kit for normalization (Sigma-Aldrich).

4.2.1.3. UPLC-UHR-ToF-MS analysis

Lipids were analyzed as previously described [68]. Briefly, lipids were separated on a Waters Acquity UPLC (Waters, Eschborn, Germany) using a Waters Cortecs C18 column (150 mm x 2.1 mm ID, 1.6 µm particle size, Waters, Eschborn Germany) and a linear gradient from 68% eluent A (40% H_2O/60% acetonitrile + 10 mM ammonium formate / 0.1% formic acid) to 97% eluent B (10% acetonitrile/90% isopropanol + 10 mM ammonium formate/0.1% formic acid). Mass spectrometric detection was performed using a Bruker maXis UHR-ToF-MS (Bruker Daltonic, Bremen Germany) in positive and negative ionization modes using data-dependent acquisition to obtain MS^1 and MS^2 information. For every ten samples, a pooled QC was injected to check the performance of the UPLC-UHR-ToF-MS system and was used for normalization.

4.2.1.4. Data processing UPLC-UHR-ToF-MS

Raw data were processed with Genedata Expressionist for MS 12.0 (Genedata AG). Preprocessing steps included noise subtraction, *m/z* recalibration, chromatographic alignment, and peak detection and grouping. Data were exported for Genedata Expressionist for MS 12.0 Analyst statistical analysis software and as .xlxs for further investigation. Maximum peak intensities were used for statistical analysis, and data were normalized on the protein content of the sample, and an intensity drift normalization based on QC samples was used to normalize for the acquisition sequence.

4.2.1.5. Lipid identification

Lipid identification was performed using the workflow described in chapter 3.6 in combination with the predicted lipid database from chapter 2.3 and wormLipidBlastR from chapter 3.5. Annotations from the different levels were integrated into a consensus annotation and manually verified. Additionally, Sirius was used for the annotation of sum formulae for lipids not covered by the annotation process [309, 313]. For annotation on the MS^1 level, predicted lipids, LipidMaps and SMID-DB. Annotation on the MS^2 level was performed with the library generated by wormLipidBlastR as well as different versions of LipidBlast [400, 416]. To

ensure high-quality annotations, all automatically generated annotations have been manually checked and filtered by trends along the RT dimension.

4.2.2. Results and Discussion

4.2.2.1. Curation of lipids from literature

In order to generate a first version of a *C. elegans* lipid database, lipids have been curated from the literature. Forty-five articles from the years 2007 to 2022 have been included in this curation. Lipids were searched in figures, tables, and supplementary information. Since lipids are reported on different levels of detail all of them have been normalized using the most recent Liebisch shorthand notation [207, 247]. In each article, different levels of identification have been achieved. For example, some articles reported sn-specificity, which typically cannot be achieved by the employed analytical methods. Such over-annotation has to be avoided not to overestimate biological outcomes. In order to allow a fair comparison between the literature, prediction, and detected lipids, all lipids have been normalized to the species level. In certain cases, this might collapse two different lipids (e.g. having different fatty acyl compositions) into a single entry. In total, 16347 lipid-literature associations have been collected, with over 2700 unique lipid species from 38 lipid classes. In several articles, different isomeric species of lipids have been potentially detected and reported. For comparison, the number of single distinct isomers has been counted per publication (Figure 20). Numbers ranged from 11 to 1444 detected lipid species (Figure 20). The highest numbers were found in Savini *et al.*, Smulan *et al.*, Liu *et al.*, Molenaars *et al.,* and Gao *et al.* [130, 163, 447, 448]. From all lipids, the total number of carbons, double bonds and oxygens have been isolated for comparison of the combinatorial space cover by the different lipid classes.

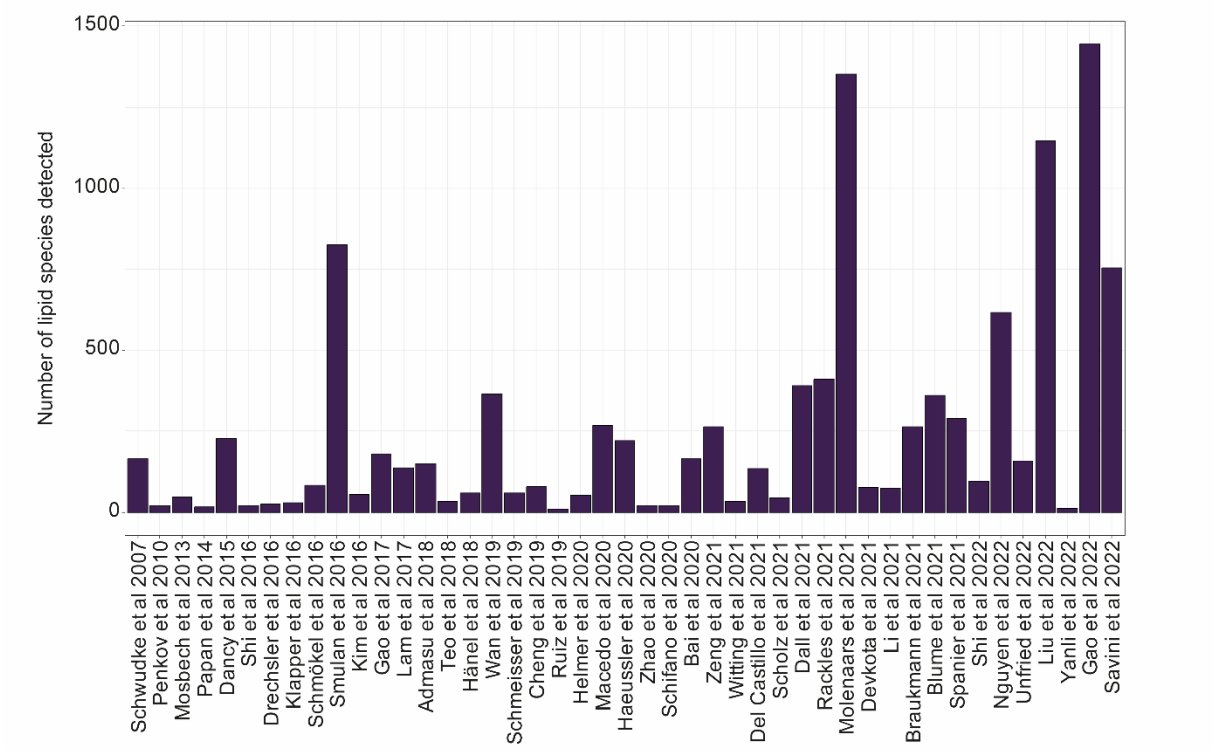


Figure 20: Number of unique detected lipid species per publication.

An extensive range of different lipids was detected. Therefore, it was checked how often lipid species were detected in the various publications. Certain and particular lipid classes have been only detected in single publications specifically dealing with them. However, single detections also included specific lipid species from classes covered in multiple publications. Numbers ranged from detection in single publications up to maximum of 26 publications. However, only a single species was detected 26 times (PC 36:5). Filtering for lipids detected in 5 publications, 715 lipid species remained. If ten publications are used for filtering, only 368 species remain. These lipid species might represent the highest abundant lipids. This included 124 TGs, 81 PEs, and 80 PCs. Differences in detected lipids can be explained by different experimental setups, which included the number of worms used for lipid extraction, the exact extraction method as well as the employed analytical setups. Furthermore, differences in genetic backgrounds and culture conditions can lead to changes in the lipidome. Though most of the experiments used *E. coli* OP50 as a food source, several studies performed RNAi, which used the *E. coli* strain HT115. It is known that nutritional differences between them can lead to changes in the *C. elegans* metabolome and lipidome.

Different mutants used in the different studies might produce different lipids, which are only detected in this condition and potentially not found in other conditions.

4.2.2.2. Overview of used data sets

Different datasets from collaborative projects have been used for the identification and in-depth investigation of lipid species in *C. elegans*. Though the culture conditions of *C. elegans* have been defined by Sidney Brenner, differences in the metabolome and lipidome can be observed based on the exact laboratory-specific protocol [4]. This effect has been examined by Spanier *et al.,* for example, studying the differences in the lipidome of wildtype N2 and *daf-2* mutants [449]. Four different laboratories produced similar samples, and lipidomics was performed centrally on all of them using the same UPLC-UHR-ToF-MS method. Though results were similar and similar “biomarkers” could be retrieved, differences in exact fold changes and p-values could be observed. This is due to the difference in food and feeding behavior of *C. elegans*. Different exact media compositions due to different used chemicals and growth conditions for feeding bacteria have been used, leading to differences in the nutritional value of *E. coli* and also the fatty acid profile, which in turn leads to differences in the *C. elegans* lipidome.

This effect is advantageous in defining the lipidome of *C. elegans* since lipids that might be missed in the analysis of samples from one laboratory might be detected in another. Here, three data sets obtained from the same UPLC-UHR-ToF-MS platform were used. Similar to the points raised above, different culture conditions in different laboratories have been used. Lipids found in all studies will form the C. elegans core lipidome, while all additional lipids are the auxiliary or pan lipidome. In the first study, Haeussler *et al.* studied changes in lipid metabolism in mitochondrial dynamics [165]. The second study by Rackles *et al.,* peroxisomal retrograde signaling was studied, while the third study focused on the role of the HLH-30 transcription factor in starvation [168, 187].

Raw data from all studies have been reprocessed using an optimized Genedata Expressionist for MS workflow, and the same annotation workflow and databases have been

used. Table 27 summarizes the overview of all datasets. Data were filtered for lipid features present in all QC samples and with an RSD < 30%. Since this optimized workflow for in-depth investigation has been developed number of lipid features and identified lipids can deviate from the values reported in the original publications.

Table 27: Overview on the three used datasets. Each dataset was reprocessed with an optimized Genedata Expressionist for MS workflow QC was performed as indicated above.

Dataset	Polarity	All	After QC
Haeussler *et al.*	(+)	8698	5390
	(-)	764	585
Rackles *et al.*	(+)	7708	5150
	(-)	861	591
Dall *et al.*	(+)	12339	6645
	(-)	3445	1738

4.2.2.3. Lipid identification

Lipids in the different datasets were identified by matching features against different lipid compound libraries on the MS^1 level, e.g. *in silico* generated *C. elegans* lipids, LipidMaps computational database, and SMID-DB and their associated MS^2 spectra against MS^2 spectral libraries derived from wormLipidBlastR or LipidBlast. All IDs were manually combined and compared in unified in consensus identifications. First, all features with an associated MS^2 spectrum were investigated. They serve as anchor point for all further investigations and their identifications from different libraries were compared against each other and manually verified based on known fragmentation pathways. In parallel Sirius and CSI:FingerID were used for verification of the formula and fragmentation pattern. In a second step, features with no MS^2 were checked. Annotation is based on *m/z* and RT has been used as a secondary filter, checking for RT trends from MS^2-verified hits. For features where no MS^2 spectra were available or that didn't match a library spectrum MS^1 data was used.

Though this way of annotation yields many false positive annotations, lipids follow a specific trend along the RT dimension. Annotations were manually filtered for entries that are not fitting these lines. First, trends for the expected main adducts were checked by plotting the *m/z* against RT for groups of lipids having the same number of carbons in all side chains. Second, all candidates not fitting this specific trendline were eliminated. In a third step, data from secondary adducts were checked for the corresponding main adduct; if it has been found, annotations were kept or otherwise removed. Several features were detected in positive and negative ionization modes. This allowed transferring annotations from one to the other ionization mode since the same chromatographic method was used. Specifically, glycerophospholipids were often detected in both modes. This improved annotation in several cases and resolved different fatty acyl chains.

Often lipids ionize as multiple adducts. The adduct grouping in Genedata Expressionist for MS allows only defining a single main adduct and secondary adducts are derived from this annotation. In positive mode, this main adduct was set to $[M+H]^+$ and in negative ionization mode to $[M-H]^-$. However, for certain lipid classes, other main adducts might be observed, e.g. for DG and TG, which ionize mainly as $[M+NH4]^+$, but also form often $[M+Na]^+$ adduct for highly abundant species. MS^1, MS^2 annotations and similarity in RT were used to manually identify secondary adducts.

This data was used as input for comparing the different datasets. This vigorous filtering enabled the highest data quality. For example, in the *hlh-30* dataset, initial annotations for DGs were reduced from 383 to 283. While only wrong MS^2-based annotations not fitting the retention time trends were removed, mostly wrong MS^1 annotations have been filtered. These represent potential false positive annotations and shows how useful the secondary filtering step using RT was. Similar results were obtained for all other lipid classes as well.

4.2.2.4. Comparison of predicted, curated, and detected lipid species

In the next step, lipids from the different datasets were compared against the predicted and curated lipids. All lipids have been normalized to the species level. By this, different lipid

isomers might be merged into single lipids species entries. The number of potential isomers are discussed later. An initial comparison was performed by plotting the covered lipid space. On the x-axis the number of carbons in all side chains and on the y-axis. If a lipid with the corresponding composition was detected, a tile is plotted at the respective position.

Figure 21 summarizes the covered space for PAs, DGs, PCs, PEs, and PS. These lipid classes are synthesized from each other and allow easy comparison. While for the prediction, always exactly the same space was covered, due to no restrictions or constraints, for the curated and detected lipids, differences between the different classes were observed. If all lipids found in the literature are allowed, a similar or larger space is covered. Upon restriction to 5 or 10 detections in individual publications this space is reduced drastically. Since not all publications used for curation performed global lipid analysis and some focused only on specific lipid classes, the filter with five publications is used. This filtering step is necessary to avoid potential “one-hit-wonders” and therefore stabilizes the entire dataset and adds some stability to the data, limiting it to only commonly detected lipid species. PCs generally showed the most significant chemical space, while PA showed the smallest. Following the biosynthetic route of these lipids, they all derive from PA. DGs are yielded either from PA by removal of the phosphate group or from other lipids, e.g. TGs, by the action of different lipases. However, larger in the covered space, the DG space of curated and detected lipids is still smaller than the predicted one. A large space is covered by PCs and PEs, and comparable spaces are covered by the literature curation and the lipids detected in the used datasets. Interestingly, the space covered by PS is very limited. Comparison of the spaces covered by shows that lipids are potentially synthesized from a limited pool of substrates producing PA and then variety in fatty acyls is added to different lipid classes by remodeling. This is also further substantiated by the limited space covered by DGs.

PGs and PIs are synthesized on different routes compared to other phospholipids, and both are derived from CDP-DGs. Comparing the covered space with, for example, PCs, a much

narrower distribution is observed. Unfortunately, PGs could not be detected in the datasets used in this chapter. However, PIs showed limited distribution in carbons ranging from 36 to 40 carbons but some variation in the number of double bonds. GC-based analysis of fatty acid analysis has already demonstrated high proportions of FA 18:0 and FA 20:5 [151]. Comparing the PIs in more detail between the used datasets, the most common species are found for 38 and 40 carbons in total. Consistent with previous GC-MS analysis, fragmentation in negative ionization mode showed all species detected, for example, in the data set from Dall et *al.* contained a saturated or monounsaturated fatty acyl in combination with a polyunsaturated fatty acyl. Only two exceptions were found (PI 18:2_20:5 and PI 20:5/20:5).

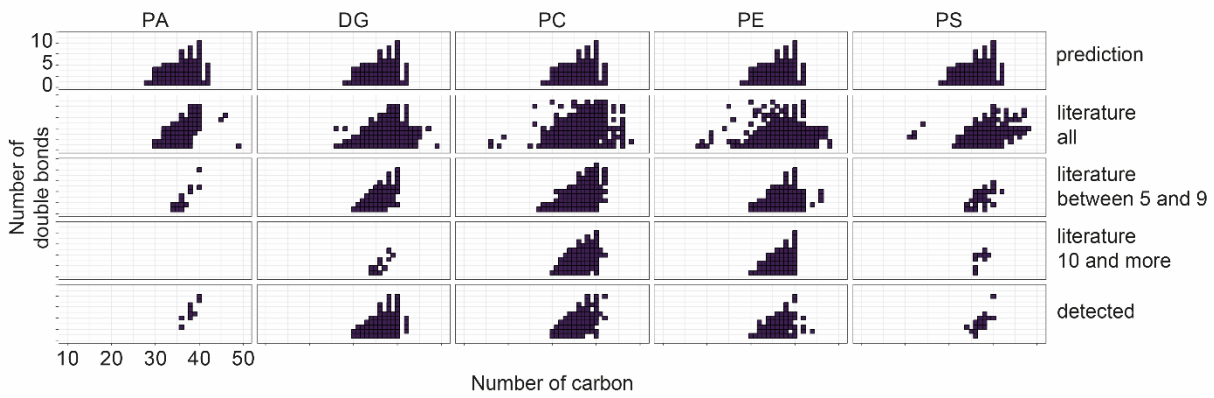


Figure 21: Lipid species covered by the different approaches to curate the C. elegans lipidome. As example the lipid classes PA, DG, PC, PE and PS are shown. These lipid classes are converted into each other and form a dense metabolic network. A tile is drawn if minimum once the lipid species with the number of carbon and number of double bonds has been predicted, curated or detected.

CLs, mitochondrial lipids, are produced on the same biosynthetic pathway as PI and PG and are also derived from CDP-DGs. CLs represent an interesting case since this lipid class is the only one that has been constrained in the prediction of lipid species. Since CLs contain four acyl chains, they can cover ample combinatorial space. CLs have not been extensively detected in the different publications. Only a few species are covered in 5 publications, and no species was found to be present 10 publications. The space covered by the CLs detected in the three different datasets is also comparably small, indicating a very narrow distribution of possible species in this lipid class. The composition of CLs has been recently studied by Oemer *et al.*, comparing CLs from different branches of the tree of life [296]. Each organism

showed a very distinct distribution. Interestingly, *C. elegans* has a very distinct composition to the CLs described by Oemer *et al.,* with CL species showing 74 carbons and more, even up to 80. Likewise, a high degree of desaturation is observed. *C. elegans* cardiolipins show a very distinct fatty acid profile and lipids annotated as cardiolipins are mirroring this composition. First, it was examined if the different detected species followed a specific trend along the RT dimension. Indeed, a clear trend was visible, indicating that all detected lipid features are probably belonging to the same lipid class; for all six putative cardiolipins MS/MS spectra were available. Second, these fragmentation patterns were analyzed. All of them showed prominent fragments, which were equal to a $[M+H-H_2O]^+$ of the corresponding DGs. Therefore, a neutral loss of the phosphorylated glycerol backbone with two attached fatty acids in combination with the NH4 occurs. Thus, “symmetric” cardiolipins would show only one specific fragment, while asymmetric ones have different fragments. The fragmentation pattern of the lipid feature annotated as CL 80:16 showed only one fragment corresponding to a composition of 40:8. In contrast, the clustered annotated as CL 80:14 showed three different peaks corresponding to the compositions of 40:8, 40:7, and 40:6. Using this information two possibilities to form CL 80:14 exist and peaks are co-eluting. Based on the fragment intensities, the symmetric CL is the major species. Fragments of all other lipid clusters were in agreement with the described fragmentation described in LipidBlast [382].

One of the missing lipid classes found in the literature but not predicted nor detected is TG-O. In contrast to TGs in which all three side chains are bound acyl, in TG-Os, one of them is bound as alkyl or alkenyl. They are derived from DG-Os potentially by the promiscuity of diacylglycerol acyl transferases, also accepting DG-O in addition to DGs. DG-Os, in contrast, are derived from alkyl phospholipids such as PC-O or PE-O. Comparing the chemical space of TGs and TG-Os, it is visible that the alkyl species cover a much smaller chemical space. This is also the case when comparing PE-O and PC-O between the prediction, curation, and detected lipids. The limited space in the TG-O is explained by the limited space of the substrates DG-O. The dataset from Dall *et al.* was inspected for potential TG-Os. Indeed *m/z*

values potentially indicate the presence of TG-Os. Unfortunately, only the MS^1 could be checked since no MS^2 data collection has been triggered due to low intensities.

Two other lipid classes were only found so far in the literature so far, PE-NMe and PE-NMe2. They have been only detected by Schwudke et *al.* [100]. This lipid class could not be identified so far in any of the datasets used. These lipids have been observed in mutants of *pmt-1* and *pmt-2*. It is not entirely clear if PMT-1 and PMT-2 act on the precursor of the headgroups (free phosphoethanolamine) or on the complex lipids. So far, the intermediates seem to be only found under these conditions. Therefore, also no pathway can be added to WormJam, and the lipid species are therefore not predicted. Likewise, maradolipids have been only found in the literature so far in three articles [80, 154, 155]. Also, cholesteryl esters have not been added to the metabolic model, though they were detected in different publications and in all three used datasets. Different lipid classes have been detected, but no predictions are available for them, e.g. lyso cardiolipins. Since the remodeling of CLs is currently missing in the prediction, no lysoCL are covered. Likewise, no remodeling of PI and, therefore, no LPIs are covered in the WormJam GSMN. Another lipid class not covered so far in the model is Bis(monoacylglycerol)phosphate (BMP), representing isomer structures to PGs. BMPs are low-abundant species and rare in most mammalian tissues [450]. They typically show enrichment in endosomal and lysosomal vesicles. They have been only detected in two publications [448, 451]. In the case of Molenaars *et al.* a normal phase separation has been employed, which allows the separation of PGs and BMPs chromatographically [448]. No biosynthetic route has been proposed in *C. elegans*, and only 17 species have been detected. It remains elusive how important this class will be in the nematode and in which tissues it might be enriched. Other lipids reported in the literature are acylcarnitine and N-Acylethanolamides, which are normally covered by methods for non-polar metabolites and are therefore only partially considered as part of the lipidome and not further discussed here, though their importance (e.g. NAEs are important signaling molecules).

One particular lipid class was described by Boland *et al.* [452]. They described for the first time phosphorylated glycosphingolipids, which are required for the mobilization of cholesterol. These molecules are derived from GlcCer and have an additional phosphoethanolamine or monomethyl phosphoethanolamine group attached to the sugar moiety. So far, this class of molecules which are named PEGC or mmPEGC, have been only identified by this group. Interestingly, this class of sphingolipids contains a hydroxy sphinganine base normally referred to as phytosphingosine, which is not detected in other sphingolipids. Since this class of lipids represents a novel derivate of sphingolipids, the datasets were manually searched for potential candidate features. No reference database for MS^2 spectra of these lipids nor their structures has been deposited. However, fragmentation is extensively explained, and reference spectra are found in the supplementary information of Boland *et al.* [452].

Table 28: Names, sum formula, and m/z of most abundant PEGC and mmPEGC according to Boland et al. [452].

Name	Formula	*m/z*
PEGC-C22	$C_{47}H_{95}N_2O_{13}P$	$[M+H]^+$ 927.6645 $[M+Na]^+$ 949.6464
PEGC-C24	$C_{49}H_{99}N_2O_{13}P$	$[M+H]^+$ 955.6958 $[M+Na]^+$ 977.6777
mmPEGC-C22	$C_{48}H_{97}N_2O_{13}P$	$[M+H]^+$ 941.6801 $[M+Na]^+$ 963.6620
mmPEGC-C23	$C_{49}H_{99}N_2O_{13}P$	$[M+H]^+$ 955.6958 $[M+Na]^+$ 977.6777
mmPEGC-C24	$C_{50}H_{101}N_2O_{13}P$	$[M+H]^+$ 969.7114 $[M+Na]^+$ 991.6933

The dataset from Dall *et al.* was first examined. Several features with fitting *m/z* values were found. However, they were generally low in intensity, and MS^2 has been only triggered for two features, potentially annotated as mmPEGC-C22 and -C24. The fragmentation of both has been described. Common fragments are *m/z* 156.042, 282.073, and 300.0842, which correspond to the “headgroup”, the monomethyl phosphoethanolamine glucose moiety. Other fragments correspond to the ceramide and vary between the species. Other species

could be only identified via their accurate *m/z*. Further variants with shorter and longer fatty acyl side chains were checked. However, when checking the trendline along the RT, it was found that all of them are forming a trendline. In total, five species of mmPEGC were putatively annotated (C21 to C25). In the datasets from Häusler et *al.* and Rackles et *al.,* C22 to C24 and C22 to C25 could be putatively annotated. This indicates that PEGC and mmPEGC are more widespread, and missing referencing in the database is potentially a main cause of missing identifications.

Three lipid classes, namely CerP, CDP-DG, and PGP, were only found in the prediction from WormJam. CerP is an important precursor for the highly potent signaling molecules sphingosine-1- phosphate [453]. CDP-DG is an essential intermediate for the synthesis of PI and PGP, PG, and CL. The lipids themselves might be present only low in concentration and are potentially unstable and, therefore, not detectable.

So far, the comparison has been carried out on the species level. However, several more lipids are present on the higher detail levels, such as molecular or structural subspecies. For example, Admasu *et al.* detected 11 different isomers for TG 55:3 [170]. Based on the theoretical predictions, 552 isomers would be possible for this sum composition on a full structural level. Other examples include species such as PC 38:5 with up to 13 isomers and 28 theoretical possible. The use of reversed-phase chromatography allows the separation of such isomers. Detailed analysis in negative ionization mode normally allows for deriving information on the bound acyl groups. However, often chromatographic separation is not perfect, and peaks might overlap. Deconvolution of peaks with different compositions is a difficult task. Underneath one chromatographic peak potential, several further isomers are found, but often only the most abundant molecular subspecies is reported instead of all detected, or only the species level is used. Using the example of PC 38:5 following compositions have been reported in different publications, which reported molecular species: 18:0_20:5, 18:1_20:4, and 18:2_20:3. These findings have been verified by the identifications in all three used datasets. However, in an MS^2 for PC 18:1_20:4 in the *hlh-30*

data, traces of other acyl groups have been detected. Comparing the number of isomeric species across different lipid classes, similar numbers were obtained between the different used datasets. The highest numbers of isomers were observed for TGs, followed by PCs, PEs, and DGs. However, the numbers are not comparable to the number of predicted isomers.

4.2.2.5. Acyl side chains detected

Besides the comparison of detected lipid species, in the next step, the acyl chains contained in the lipids were compared. Since not all publications report acyl chain species, Savini *et al.*, Smulan *et al.,* and Lam *et al.* were selected as examples since they reported lipids with fatty acyl compositions [163, 171, 447]. Detection in negative ionizatoin mode allowed to identify acyl side chains in glycerophospholipids, and positive ionization mode neutral losses specific to the different acyl side chains have been used. Fatty acid analysis by GC-FID or GC-MS enabled the detailed description of fatty acyls in *C. elegans* lipids and served as a comparison. While analysis with LC-MS/MS does not allow to derive as detailed description regarding the double bond position and stereochemistry as GC-MS, a comparison on the "species" level of the acyl chains is possible. Sometimes multiple mapping between isomers detected by GC-FID or -MS and acyl side chains exist. For example, 19:1 can be mapped to FA 19:1(10Z) or FA 19:0;[9-10cy3:0].

Interestingly, much more acyl side chains than expected have been detected, including several unusual ones, which included, for example, 16:1, 16:2: 16:3, 16:4 or 19:2, 19:3, 19:5. These unusual side chains have been detected in all of the here used publications, but so far not in the datasets. A recent metabolomics study focused on *hacl-1* and different acyl-based metabolites [454]. Several acyl groups also found in lipids were detected in this study as well. Checking the three generated datasets, fragments corresponding to FA 19:3 was searched in the negative ionization mode data. Indeed, traces of fragment *m/z* values corresponding to this acyl group could be detected from complex lipids. While all major acyl groups which have been previously also detected by GC-FID or -MS are identified in all selected

publications and datasets, the more unusual acyl groups represent minor species. It remains to be elucidated if these acyl groups are of functional relevance or represent only artifacts arising from the promiscuity of enzymes. Furthermore, the source of the fatty acyls is unknown. HACL-1 is required from α-oxidation of fatty acids, which is involved in the degradation of unusual fatty acids. Brock *et al.* identified alternative, unique fatty acids in *fat-6;fat-7* mutants, which included 16:2 and 16:3 produced by FAT-2 and FAT-3 [258]. A further potential source might be the β-oxidation of longer polyunsaturated fatty acids and the possible release of the acyl groups from Acyl-CoAs. Since most of the unusual fatty acids are only minor species, they are only possible side products. Brock *et al.* performed GC-MS analysis of total fatty acids; therefore, it is still unclear how these unusual fatty acids distribute across different lipid classes. Interestingly, when checking lipid classes in which these fatty acids were annotated, it was mostly PCs. These fatty acids might be also present in other lipid classes. However, PCs are the most abundant lipids.

4.2.2.6. New detected lipid classes

Since spectra matching has also been performed against different external libraries, not only *in-silico* spectra for predicted species in *C. elegans,* several more lipids could be identified, including new lipid classes previously not detected in *C. elegans*. One example is fatty acid esters of hydroxy fatty acids (FAHFAs). FAHFAs have been initially described in adipose tissue, and their concentration correlates with insulin sensitivity [455]. Typically, these lipids contain a long chain hydroxy fatty acid, e.g. C16 or C18 fatty acids, and to the hydroxy group a second fatty acid is esterified. However, the FAHFAs detected in *C. elegans* showed different compositions and contained a C9:0;O fatty acid and have been annotated using LipidBlast [456]. LipidBlast uses various fatty acids, including some unusual ones, to generate possible lipid species. First, spectra putatively annotated as FAHFAs were additionally inspected using Sirius for calculation of the sum formula for precursors and fragments. In the case of all putative C9:0;O containing FAHFAs, the correct and expected sum formulas were calculated for the precursor and fragments. Next, retention times were examined. All detected FAHFAs elute between 10 to 15 minutes. This area is typically

covered by highly unsaturated glycerophospholipids such as PC 40:10, etc. Comparing logP values as predicted by MarvinSketch for representative structures, this elution range is fitting. For most of the detected FAHFAs, multiple chromatographic peaks were found. Often three peaks eluting very close to each other with no baseline separation were observed. These peaks might represent different isomers of the detected FAHFAs, and all showed very similar fragmentation spectra with only differences found in the intensity ratios. Recently, peaks potentially annotated as FAHFAs in the non-targeted analysis have been described to be artifacts of fatty acid dimers and do not represent true hits [457]. The most important indicator for correctly identifying a FAHFA is the presence of the $[M-H]^-$ and $[M-H_2O-H]^-$ fragments of the hydroxylated acyl chain. This was the case for all detected FAHFAs (*m/z* 155.1078 and *m/z* 173.1183).

Additionally, no evidence of the presence of the individual free fatty acids detected at the specific retention time range was found. The separation of different isomers of FAHFAs has been described but requires long gradients and optimized conditions [458]. It has been shown that FAHFA isomers elute according to the position of the hydroxy group and the esterified fatty acid. In the case of palmitic acid ester of hydroxy stearic acid (PAHSA), the elution order was 12-, 11-, 10-, 9-, 6-, 5-, and 4-PAHSA. Additionally, ratios of different fragments were different, e.g. the ratio of *m/z* 255.2329 and *m/z* 299.2592, corresponding to the esterified palmitic acid and intact hydroxystearic acid fragment increased with the hydroxy position moving closer to the carboxy group. Plotting the ratio of the respective fragments against the retention time, first, an increase and then a decrease was observed. Since no reference standards for the detected FAHFAs are, available further identification at higher MSI levels was not possible. However, it is intriguing that the hydroxy fatty acid in *C. elegans* FAHFAs is C9:0;O. The most abundant ascarosides in WT *C. elegans* are ascr#3 and ascr#10, which are also C9 species that have an α,ß-unsaturation, or are completely saturated. No FAHFAs containing a C9:1;O fatty acid could be found. However, searching for further potential FAHFAs by searching for spectra containing *m/z* 173.1183, several additional candidates were detected. These lipids are eluting in the range of TGs towards the

end of the gradient but are only detected in negative ionization mode. Manual inspection of their fragmentation pattern showed the loss of an additional C9:0;O fatty acid from the precursor followed by a similar fragmentation to FAHFAs. Based on the increased RT and the additional loss of C9:0;O a structure in which an additional hydroxy fatty acid is esterified is suggested. Sirius predicted the correct formula for the precursor and the fragments of these peaks. For each formula multiple isomers with different intensity ratios of fragments exist, which also suggests the presence of multiple species with different positions of the hydroxyl group.

Besides these FAHFA, several further unknown lipids exist. Only a small part of the detected lipids could be annotated. For example, very long ascarosides have been putatively annotated on the MS^1 level. Furthermore, some evidence for more acyl-based lipids, such as N-acyl amino acids, have been found but have not been further followed up.

4.2.2.7. Database of lipids

To enlarge the database of lipids found in *C. elegans,* the list of the curate, predicted, and detected lipids has been combined. This combined database has been built on the species level. The number of potential lipids covered will increase even further if more detailed levels are incorporated. From the initial 1690 lipids predicted the list was enlarged to 3284 lipids in total. This included new lipids from classes that were not included in the prediction and new species. For example, TGs increased from 222 to 422; PCs increased from 95 to 187; PEs from 96 to 176; and PE-O from 52 to 89.

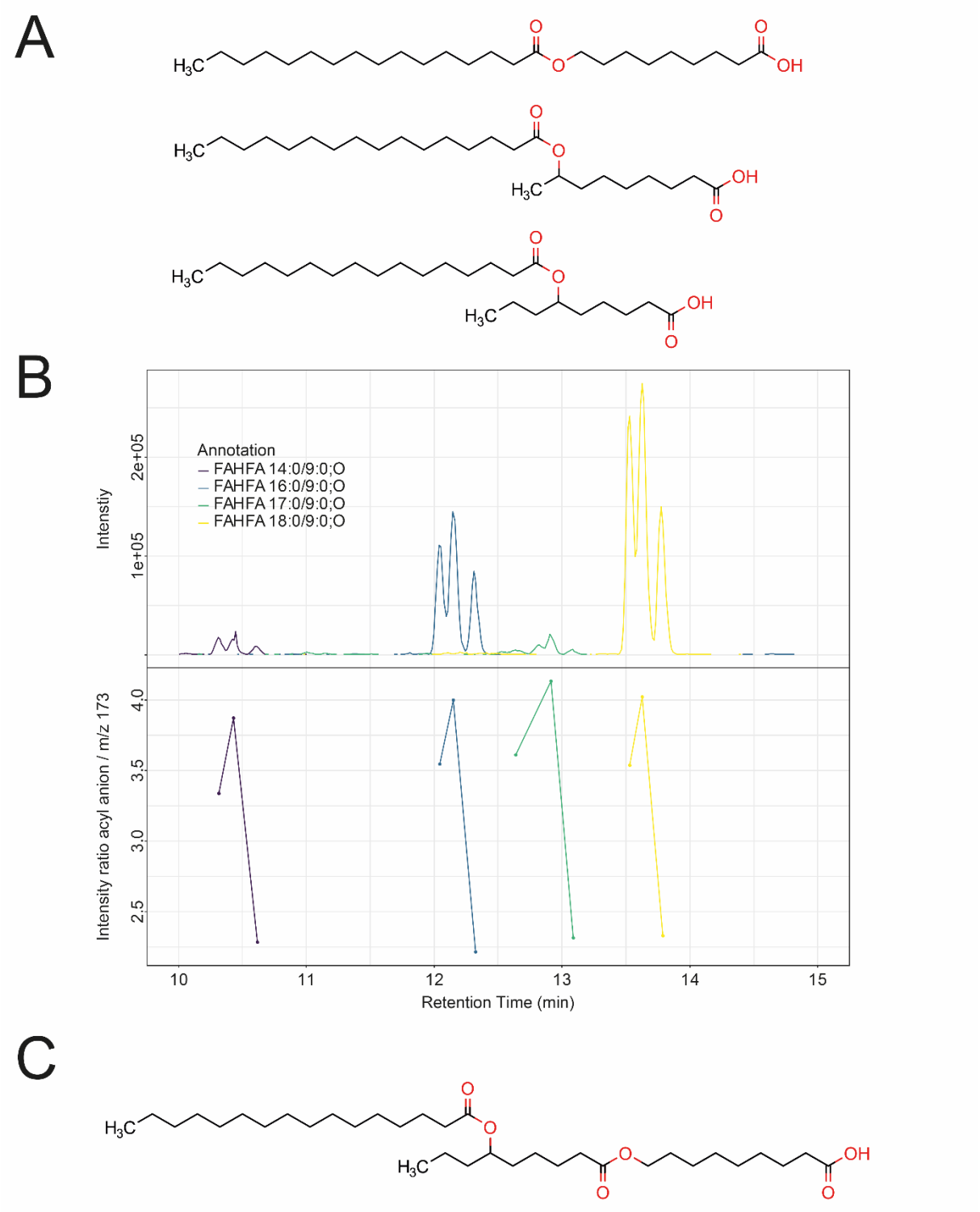


Figure 22: (A) Putative structures for C. elegans FAHFAs. The isomeric structures differ in the position of the hydroxyl group on the C9 fatty acid. (B) Extracted ion chromatograms for different FAHFA species indicate several isomeric species. The ratio of the acyl chain anion over m/z 173.1172 (FA 9:0;O) is different for each isomer indication different positions of the hydroxyl group. (C) putative structure of FAHFAs with an additional hydroxy fatty acid esterified.

For all lipids in the combined database, molecular formulae are available, which enables the calculation of theoretical masses and *m/z* values from them for annotation of lipids in future

experiments. Furthermore, for all cases in which compositions are available, they can be used for the prediction of MS^2 spectra. In total, 2825 unique chemical formulae are contained. Interestingly, the different lipid classes often also contain oxidized species for certain lipids. These oxidized lipids play important roles in messaging and are potentially present only in low amounts compared to nascent species. Oxidation can occur during biological processes or as an artifact of improper sample preparation. A more detailed characterization of these species will be required in the future.

4.2.2.8. Biological results

Many identified lipids also showed significant differences between the conditions studied in the specific experimental setups. All used datasets deal with mitochondria and peroxisomes as organelles for energy production. First mitochondrial dynamics have been studied in *fzo-1*/MFN2 loss of function mutants. FZO-1 is involved in mitochondrial fusion, and loss of function leads to mitochondrial fragmentation, which induces mitochondrial unfolded proteins response (UPR^{mt}). Investigating this mutation using lipid profiling, it was found that mostly TGs are affected. Within TGs, a clear separation of different species and their associated changes were found. While TGs species with a lower number of carbons and double bonds are downregulated, species with a higher number of carbons and double bonds are upregulated. In parallel to *fzo-1* also *drp-1* loss of function mutants were examined. DRP-1 is involved in mitochondrial fission. Also, in *drp-1* mutants, some TGs are changing, but no clear trend similar to *fzo-1* could be identified. As additional control *spg-7*, encoding for a mitochondrial metalloprotease required for the correct mitochondrial function was used. *spg-7* mutant worm showed down-regulation of several TG species. Using the sum of all features identified as TGs as a proxy for the lipid droplet content no significant differences between wildtype and *fzo-1* loss-of-function were found, while *drp-1* showed a slight but significant increase and *spg-7* a considerable decrease. A genome-wide RNAi screen for genes suppressing the induction of UPR^{mt} in *fzo-1* mutants has been conducted. *vps-4* and *cogc-2* have been identified.

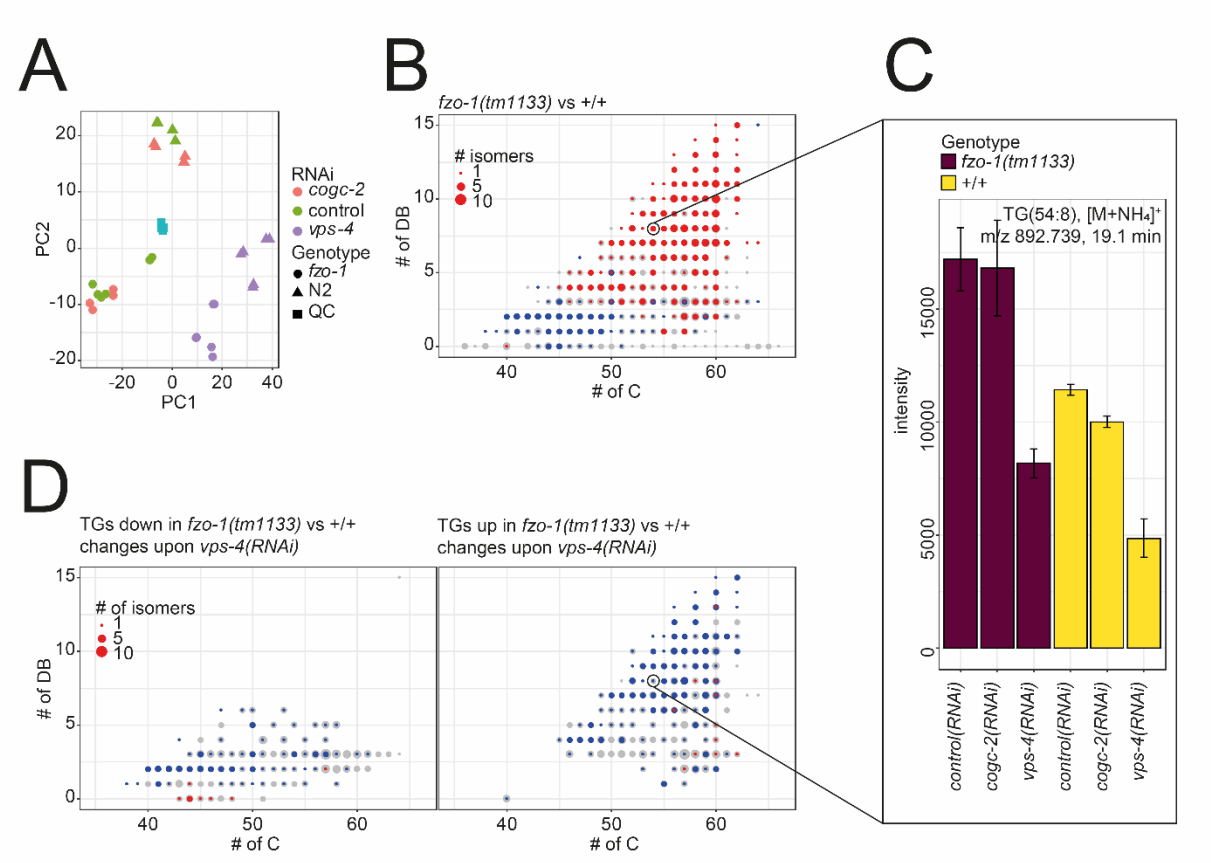


Figure 23: (A) PCA of lipidomics data comparing different RNAi in different genetic backgrounds (N2 and fzo-1(tm1133)). vps-4 RNAi treated animals show profound differences by forming a separate cluster. (B) TGs species covered and changed between N2 and fzo-1(tm1133) worms. Generally, lipids with lower number of carbons and double bonds are downregulated, while longer and more unsaturated species show and increase (C) Bar plot of TG 54:8 as example of one changed lipid. (D) lipids changed in the additional treatment with vps-4 RNAi.

The results indicate mitochondrial dynamics in *fzo-1* loss of function mutants might affect, to a certain extent, mitochondria-associated with lipid droplets, the storage organelles for TGs. These peri-droplet mitochondria might have unique morphology and function to support the synthesis of TGs and generation of lipid droplets. Indeed, Benador *et al.* have isolated such peri-droplet mitochondria from mouse brown adipose tissue (BAT) [459]. Compared to cytoplasmic mitochondria, these mitochondria have increased pyruvate oxidation, electron transport, and ATP synthesis but reduced β-oxidation capacities. Furthermore, the protein composition is different. All results suggest that peri-droplet mitochondria are a separated population supporting TG biosynthesis and lipid droplet generation. *fzo-1* mutation might especially affect this subpopulation of mitochondria. CLs are important lipids for the correct function of mitochondria.

As an important organelle for energy production, mitochondria are also crucial for the response to changes in nutritional conditions. The dataset from Dall *et al.* studied the role of the HLH-30/TFEB transcription factor during starvation. Multiple time points have been analyzed from fed and starved WT worms or *hlh-30* loss of function mutants. While in WT worms, different changes along the different timepoints of starvation were found, most of these changes were absent in *hlh-30* mutants. This included features annotated as CLs. Interestingly, in WT, these peaks were significantly upregulated across all examined time points except for 16 hours, while in the *hlh-30* mutants, they should consider a downregulation trend, which was not significant at any time point. Using published RNAseq data, it was found that *hlh-30* has no known target genes in cardiolipin metabolism, although in influences several genes of mitochondrial β-oxidation. Therefore, the mechanism by which HLH-30 is influencing CL levels remains elusive and further experiments are required. Upon starvation, a typical response is the mobilization of storage fats for energy production. Only a few TGs were identified to change significantly differently in the WT worms. It seems that 16 hours of starvation were not enough, and this change is only found at later time points. The increase in CLs might represent a first line of changes to cope with starvation, changing mitochondria morphology and their energy production capacity. In further experiments, it was shown that *hlh-30* mutants switch from mitochondrial to peroxisomal β-oxidation [187]. Additional knockdown of *prx-5*, required for biogenesis of peroxisomes, renders *hlh-30* mutants even more sensitive to starvation.

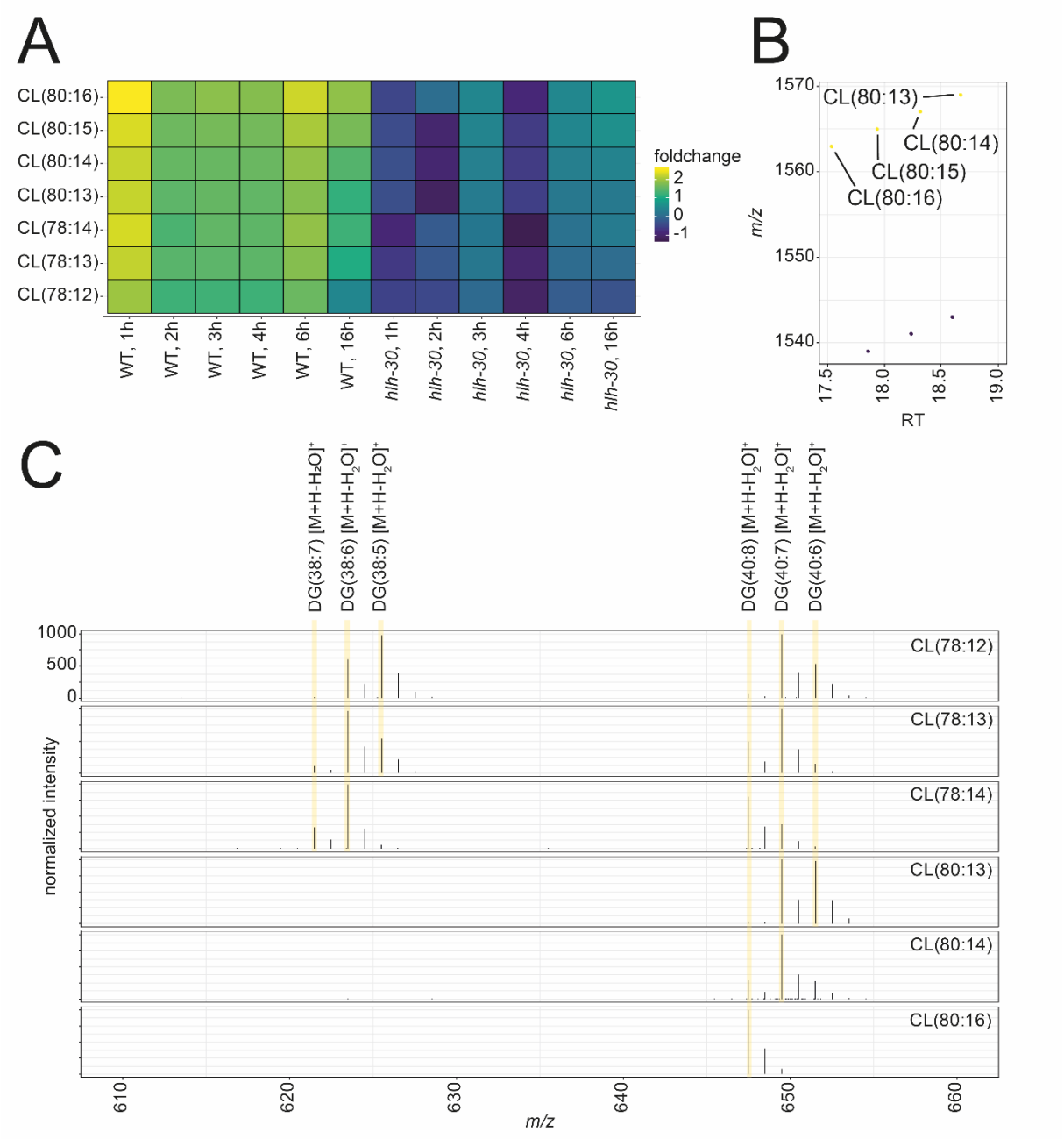


Figure 24: (A) Changes in CL species across 16 hours of starvation in WT and hlh-30 worms. While WT worms show an increase when comparing fed against starved worms with an upregulation of CLs, now changes were found in hlh-30 mutants. Changes in hlh-30 dataset (B) CL species were identified by m/z and RT as well as MS², if available. Trendlines along the RT dimension were used for filter false positive annotations. (C) Fragmentation spectra of selected CL species show that different possibilities for CL are hidden below a single chromatographic peak.

The peroxisomal matrix import receptor *prx-5* is required for the protein import into peroxisomes and their biogenesis. The dataset from Rackles *et al.* studied the effect of the knockdown of *prx-5*. Lipid profiling showed several significantly changed lipids, including several TGs. Specifically, several TGs with a higher number of carbons have been found to be upregulated. This is consistent with the role of peroxisomes in the breakdown of long-

chain fatty acids. Upon a defective or reduced number of peroxisomes, they are no longer degraded and potentially sequestered into TGs. Furthermore, peroxisomes are required for the biosynthesis of ether lipids. In line with this function, all detected PE-Os, the major ether lipid class in *C. elegans*, were found to be downregulated.

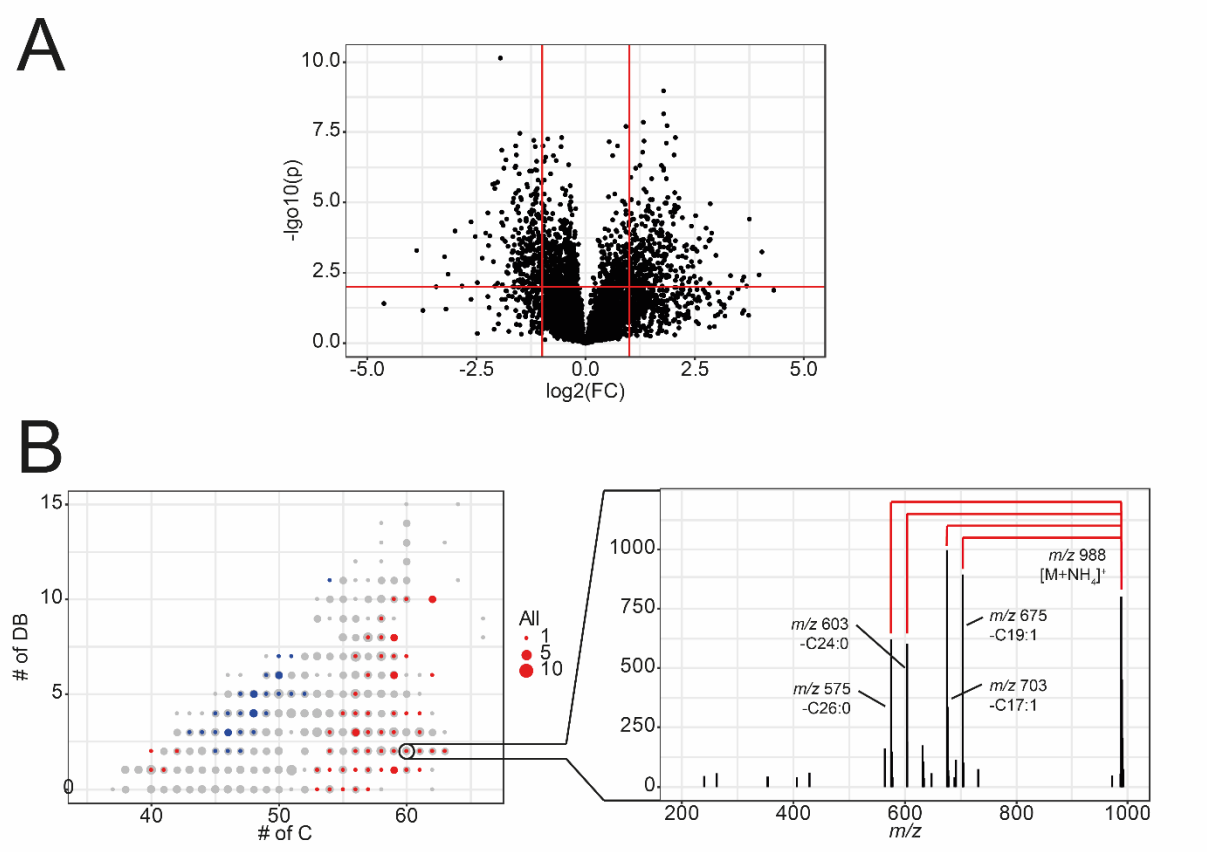


Figure 25: (A) Volcano plot comparing lipid features between N2 and prx-5 worms. (B) Different TG species were upregulated in prx-5 worms, which contain long saturated fatty acyls. The selected MS^2 showed the neutral losses for the acyls FA 24:0 and FA 26:0.

4.2.3. Discussion

Lipidomics is a fast-growing application in *C. elegans* research, and several articles have been published in this field. However, a central database containing the *C. elegans* lipidome is still missing. A first attempt was made to generate such a database by comparing lipids detected in the literature, predicted by the tools presented in chapter 2.3 and identified in own measurements. Several interesting observations have been made. One of the most important is that the prediction of lipid species needs further adjustment. In the case of several lipid classes, the covered lipid space has been vastly overpredicted, with many species that might not be present in the nematode, e.g. in the case of PIs or PSs. Here the

prediction needs to be further fine-tuned by the incorporation of rules for constraining potential substrates. However, these can only be established if the enzymes are studied in more detail.

Furthermore, although a certain substrate specificity might be established, other substrates might be simply accepted due to enzyme promiscuity. A potential way in the future might be to associate a numerical score with the prediction, reflecting the likelihood of producing this lipid. This score can include the abundance of the used acyl group and the potential promiscuity of enzymes and incorporates the abundance of other precursor lipids. The performed comparison has shown that already the comparison on a qualitative level between the prediction, literature curation, and measurement can help to identify gaps.

The curation of the literature cannot be regarded as perfect. First, the different publications have used many different “flavors” of lipid shorthand notations. This requires normalization of them to a common shorthand nomenclature, which can be performed by tools such as `rGoslin` [293, 294]. The comparison here has been performed based on the species level. In regard to the different publications, this seemed to be the fairest way of comparison, since this was the level at which most publication reported their results. Second, many different analytical approaches have been used, which do not allow a direct comparison of obtained results. Dependent on the chromatographic separation and used MS detection, different depths of lipid annotation can be achieved. In negative ionization mode, for example, the molecular subspecies for phospholipids can be determined, while in positive ionization mode this is possible for DGs or TGs. Class-wise separation such as HILIC allows determining the lipid class on typical elution patterns, while RP-based separation allows separating more isomeric species. Both have been employed for *C. elegans,* and no systematic differences in reporting have been found in the literature. Third, in many cases the data was curated from figures, tables, and supplementary information, which often was not a table with the raw data. It is important to link the data to its origin to allow the judging of the reader on the identifications, e.g. by submission to different repositories such as MetaboLights,

Metabolomics Workbench, or GNPS [54, 300, 437]. Several lipids in the publication might be detected multiple times, e.g., in both positive and negative ionization modes. If no retention times are reported matching between them cannot be performed. Raw data would allow to compare results and clean annotations.

Using selected examples from different publications which reported molecular subspecies, a comparison of detected acyl chains was performed. This comparison was mostly based on phospholipid annotations since the composition could often be determined in this case. Several unusual acyl groups have been detected, not only in the literature but also in the own measurements. These acyl groups were not included in the prediction and might represent some minor species that are only present in low concentrations. Generally, most of the measurements in the literature have been carried out in a semiquantitative manner, only reporting peak intensities, integrals, or similar. An important step in the future will be the quantitative analysis of lipids. This would not only allow comparison on a qualitative level but could ultimately lead to the integration of multiple datasets and better comparison of results. This would also show the percentage of lipids with unusual side chains. However, to develop targeted methods for exact quantification of lipid species, first, a list of lipids is required. The database compiled here will be the first step.

The lipidome, similar to the metabolome, is very condition dependent and, therefore will vary between different laboratories. A recent study by Spanier *et al.* has shown although general trends are conserved between different laboratories and their different culture conditions, exact fold changes and recovery of “biomarker” depends on the laboratory [449]. The exact composition of the lipidome will be dependent on the culture conditions and the bacterial food supplied. Based on GC-MS analysis, differences in the fatty acid composition of *C. elegans* from different bacterial strains have been shown [147]. To allow better comparison in the future between different *C. elegans* lipidomics studies it might be advantageous to also report the lipidome of the food bacteria and their fatty acid composition. Another level of quality control is the use of reference materials. For human studies, the NIST SRM 1950

(Metabolites in Frozen Human Plasma) is widely used. Bowden *et al.* compared results from different laboratories performing lipidomics on this SRM [460]. It has been shown that the inclusion of shared reference materials helps in the harmonization of MS-based lipidomics [461]. Reference material for *C. elegans* has been developed by Gouveia *et al.* [462]. The long-term stability of this material is ensured by the pooling of multiple reference samples from different batches on a rolling average. The inclusion of this material can facilitate the comparability between different studies in the future.

At the current stage, lipidomics is applied to a limited number of conditions. In the case of the dataset from Häussler *et al.*, lipidomics was applied to the final conditions identified after the initial RNAi screen. Integration of lipidomics directly in this screen might have revealed additional interesting candidates which would not have been identified otherwise. Lipidomics in this regard can be seen as a multiparametric read-out, while typical screens using fluorescent reporters can only probe for a handful.

In the future, data can be organized in a lipid atlas as an online resource. The use of common names can link data between the different studies. This would also include the studied conditions to allow more elaborate data mining. Different levels of lipid annotations can be bridged by an ontology system as used by ChEBI or SwissLipids. Such an atlas would represent an important resource for all *C. elegans* scientists aiming to understand lipid metabolism better. The current numbers of unique lipid species per lipid class are summarized in Table 29.

Table 29: Number of unique lipids species per lipid class in the combined database.

Lipid Class	Number of species
BMP	17
CAR	82
CDPDAG	95
Cer	106

Lipid Class	Number of species
CerP	38
CL	165
CoA	53
DG	140
DG-O	95
FA	71
Hex2Cer	2
HexCer	76
LCL	7
LPA	27
LPC	56
LPC-O	38
LPE	43
LPE-O	18
LPG	16
LPI	12
LPS	16
LSM	2
LysoMar	16
Mar	25
MG	57
MG-O	2
NAE	22

Lipid Class	Number of species
PA	111
PA-O	39
PC	187
PC-O	83
PE	176
PE-O	89
PE-NMe	41
PE-NMe2	41
PG	101
PG-O	2
PGP	95
PI	102
PI-O	1
PIP	2
PS	134
PS-O	6
CE	30
SM	133
SPB	39
SPBP	4
TG	422
TG-O	140

4.3. Conclusion

So far, the *C. elegans* lipidome remains to be elusive. Based on predicted, curated, and measured lipid species, a first blueprint of this lipidome has been constructed, which includes several novel lipids, which have been only detected in single publications. The data-driven reconstruction of the lipidome represents an interesting approach and can generate living data, which can be refined at any time. However, at the current stage, this is all based on manual effort since the raw data is not available in public repositories for most studies. If correct annotations are supplied in repositories, they could be automatically retrieved and checked. For example, annotations can be checked for plausibility regarding previous publications, and spectra of unknown lipids can be compared against others to prioritize potential features for identification.

- This page is left intentionally blank -

5. Novel analytical concepts/approaches

Chapter-related publications:

Tandem HILIC-RP liquid chromatography for increased polarity coverage in food analysis

Hemmler D., SS. Heinzmann, K. Wöhr, P. Schmitt-Kopplin, M. Witting

Electrophoresis. 2018 Jul;39(13):1645-1653. doi: 10.1002/elps.201800038

The presented tandem-LC setup was developed by Daniel Hemmler and myself and further optimized by the jointly supervised master thesis of Katharina Wöhr. All work has been carried out under my supervision.

5.1. Introduction

5.1.1. Metabolome coverage and LC column coupling approaches

Successful application of metabolomics relies to a certain extent on the number of metabolites detected and the metabolic pathways covered. However, metabolites span an extensive range of polarity and concentrations (over several orders of magnitude). Therefore, no single analytical method can currently cover all metabolites in a single run. To overcome the coverage issue, multiple analytical methods must be combined, e.g. HILIC and RP separation of metabolites. By this, the covered polarity range is greatly enhanced. However, multiple data tables from different methods are generated, and integration for statistical analysis is time-consuming and tedious. In a typical chromatographic setup, only retained substances are analyzed, while metabolites that elute in or close to the void volume are removed from further analysis. These are the typical metabolites that are analyzed in the complementary chromatographic separation, e.g. metabolites from the void volume in an RP separation are retained on a HILIC column. However, several metabolites can be retained under RP and HILIC conditions. In the case of known metabolites merging of multiple data tables can be performed, e.g. selecting the analysis mode yielding the best results for a particular metabolite [373]. However, unknown metabolites can often not be directly matched, or several matching possibilities might exist due to the presence of isomeric structures with the same *m/z* in the different separation modes. If MS^2 data are available, features can be matched based on these. Lastly, analysis on multiple analytical setups requires a more significant amount of biological starting material, which might be unavailable in the case of sample-limited studies, such as large-scale genetic screens.

Furthermore, it requires analysis more time due to multiple injections. Combining analytical methods that can cover a more extensive range of metabolites represents an excellent solution to overcome this obstacle. This includes novel analytical setups such as parallel LC, tandem LC or multidimensional separation techniques. Analysis using multiple chromatographic methods can be sped up by the use of parallel analysis on multiple LC-MS systems. These systems either operate in positive and negative ionization modes using the

same or different chromatographic methods. Often, a significant time frame of the analytical sequence is dedicated to column re-equilibration in which no useful data is collected. In order to increase the idle time of an MS, injections and analyses can be intercalated between different LC systems coupled to the same MS so that during the re-equilibration of one system, another or the same sample is measured on another system. This setup is called parallel LC. Klavins *et al.* developed a single injection method that allows parallel analysis of hydrophilic and hydrophobic metabolites [463]. Use of a ten-port-2-position valve and two pumps allowed to connect an RP and a HILIC column to a single injection loop. After loading samples, both methods can run individually through individual pumps. An additional valve was used for selecting the effluent which should be directed to the MS. Parallel LC showed an excellent analytical performance comparable to the individual methods. A similar setup has been realized for the parallel analysis of metabolites and lipids from NIST SRM 1950 [464]. In both applications, one column can be re-equilibrated while the other is used for analysis, saving valuable measurement time.

However, in parallel LC, the only parameter optimized for is the analysis time, not the amount of sample used. In contrast, tandem LC setups combine two orthogonal separation columns in a single chromatographic setup allowing analysis in both separation modes from a single injection. Typically, an RP and a HILIC column are coupled in series with a T-piece between the columns, via which a second pump delivers solvents for trapping and separation on the second column. Chalcraft and McCarry coupled together different HILIC and RP columns and compared the number of detected features from a pooled mouse serum sample [465]. A combination of an RP-Amide with a ZIC-HILIC yielded the highest number of metabolite features. Wang *et al.* described a fully automatic HILIC-RP switching in combination with data-dependent fragmentation on a linear ion trap [466]. Urine samples were analyzed, and in total, 5686 polar and 1808 non-polar features were detected. In 2013 Greco *et al.* proposed a serial coupling of RP and HILIC for the simultaneous analysis of polar and non-polar phenols in wine [467]. A short C18 column (50 x 3.0 mm) was coupled serially to a zwitterionic HILIC column via a T-piece. The first pump delivering the solvents through the

complete system experiences a higher pressure due to the serial coupling; therefore, a superficially porous material with a larger particle diameter was used lower the back pressure while maintaining high efficiency in separation. Different phenols from wine spanning a large polarity range could be analyzed. The system was also tested for robustness, and results indicate that it is robust and reliable over one year, with > 1100 injections performed [468]. Retention time shifts of phenols were <30 s. A similar setup was applied by Haggarty *et al.,* analyzing metabolites from beer [469]. Standards of different polarities were used to evaluate the performance of the system. RSDs were comparable to individual column separations.

A significant disadvantage of this particular setup is that both columns are never independent of each other, and both effluents are directed through the second column (HILIC in most cases). This restricts the number of employed separation systems, and the compatibility of solvents and additives is an important issue. Furthermore, HILIC columns require a long re-equilibration time, and in the represented setup, they can only be re-equilibrated after the entire analytical run has been completed. Tandem-LC follows the same idea of the serial coupling of columns using the metabolites that elute in the void volume and trap them on a second orthogonal separation chemistry. The setup employs a first trapping column on which the portioning into hydrophilic and hydrophobic metabolites is performed. Different setups are possible. In the case of an RP column, hydrophilic metabolites are eluting in the void volume and are diluted with ACN to be trapped on a following HILIC column. The difference is that a switching valve isolates the RP trapping column from the HILIC column after the trapping phase and makes independent separations possible. This setup was proposed by Pyke *et al.* and combined RP with ANP separation [470]. Using their method, they were able to detect from 1212 features in human urine to 4248 in mouse tissue extract over a large polarity range. The tandem-LC configuration was in good agreement with the separation on the individual columns showing that combined analysis is possible. Lv *et al.* combined three different columns for the comprehensive coverage analysis of metabolites and lipids from different samples using pseudo targeted detection on the MS^2 level [471].

Lastly, 2D-LC exist as possibility to increase peak capacities and coverage of the metabolome and lipidome. 2D-LC can used in different modes. First, heart-cutting 2D-LC uses a defined window in the first dimension, which is then transferred to the second dimension for analysis with an orthogonal separation method. One example was performed by Helmer *et al.* for the analysis of cardiolipins and their oxidation products in *C. elegans* [295]. Cardiolipins were separated from other lipid classes using a HILIC separation in the first dimension and then transferred to the second dimension using a RP column for separation of individual cardiolipin species. If multiple windows are transferred the method is called multiple heart cutting. Scholz *et al.* used such a system for the profiling of sphingolipids in *C. elegans* [472]. Lastly, comprehensive 2D-LC collects consecutive fractions of the first dimension and transfers them to the second dimension. High peak capacities can be achieved, especially of the two dimensions are orthogonal to each other, but data analysis is much more complicated. Therefore this method is only used by specific laboratories [473].

5.1.2. Miniaturization and throughput

Besides the increase in metabolite coverage, the required amount of biological start material is also an important factor. To enable large-scale metabolomics and lipidomics experiments with *C. elegans,* the amount of biological starting material required needs to be reduced, which might need to be reduced down to several hundreds or only tens of *C. elegans*. The required amount of worms has already been reduced for certain analyses; for example Witting *et al.* have used 5000 worms in their publication from 2014, while currently only 500 are required [68, 168, 187]. The final endpoint of this development would be the analysis of single worms. *C. elegans* show a certain degree of heterogeneity, although they are typically referred to as isogenic, which in theory should result in a uniform outcome. Under normal “population-based” metabolomics and lipidomics approaches, this heterogeneity is averaged to a population mean by pooling several hundreds to thousands of worms into a single sample. The smaller the population will get, the more the individuality of the worm will play a role.

Therefore, an important question is if the response of individual worms can be compared. The most prominent example of heterogeneity in *C. elegans* biology is survival. A typically survival curve shows a sigmoid shape, with a few induvial still alive while others already ceased. In order to be able to study differences throughout aging and even extremely long-lived individuals, approaches that can analyze small amounts of biological material are required. A simple calculation expresses this need. For a standard metabolomics or lipidomics experiment, 500 worms are required as a minimal amount. If a time point at which only 10% of worms are alive shall be analyzed, 5000 worms, which equals ten standard NGM plates, is required for one biological replicate. Worms are scattered across these 10 plates. Simple washing of the plates is not possible because also dead worms would be washed off the plates and contaminate the sample. Methods to separate alive from dead worms are possible but require time and potentially alter the metabolome. Picking 500 worms again requires a lot of time. This demonstrates the strong need for methods with high sensitivity or miniaturized methods in *C. elegans* research, not only metabolomics and lipidomics but also other technologies. The first examples showed that analysis of a low amount of worms is possible. For example, DI-FT-ICR-MS analysis of *C. elegans* samples by Witting *et al.* first hints that single *C. elegans* metabolomics might be possible [185]. Extracts of 1000 worms were required to be diluted 50 times to yield a concentration of 20 worms / mL. From this diluted extract, only a few µL were required for the actual analysis. Furthermore, Bensaddek *et al.* described a microproteomics approach for the analysis of single *C. elegans* [188]. Likewise, Wong *et al.* performed an analysis of single *C. elegans* using NMR [94]. However, a long analysis time was required, and only a handful of metabolites could be measured.

5.2. TandemLC for increased metabolite coverage in a single method

Several presented tandem-LC setups combined their effluent after separation via an additional T-piece. To not mix effluents flow from the respectively not used separation mode is switched off. Similar to the serial column setup, columns can only be re-equilibrated after the complete run is over, which represents a significant drawback. For the realization of a tandem-LC setup for the analysis of *C. elegans* sample, the setup of Pyke *et al.* was adapted but included an additional valve post column, which allows selecting the effluent which is directed to the MS, while to other is directed to the waste [470]. The basic setup is shown in Figure 26. Metabolite extracts are first injected into a short column used as a trap, and hydrophilic and hydrophobic metabolites are separated. Hydrophilic metabolites not retained elute in the void volume are diluted with ACN mixed via a T-piece and trapped on the head of the HILIC column. After the trapping phase, the first valve is switched and isolates the HILIC column from the RP separation system and switches the trap column with the RP separation column in series. First, the HILIC separation is performed by running a suitable gradient. After this separation is done, the second valve is switched, and the effluent from the HILIC part is directed to the waste, while the RP part is connected to the MS. While RP separation on the mid- to non-polar metabolites is performed, the HILIC column is re-equilibrated.

The metabolome of *C. elegans* is complex, as shown in Chapter 2. No single chromatographic method is able to cover the entire polarity range. So far, the most employed chromatographic method for the analysis is RP, and only a few applications of HILIC have been published, e.g. the use of ZIC-cHILIC column by Molenaars *et al.* or others [51, 52, 448]. In order to analyze the complete metabolic network, a combination of different chromatographic methods is required. However, integration from multiple methods is complicated, especially for unknowns. The combined analysis of polar and non-polar metabolites is of great interest for *C. elegans* metabolomic investigations. The tandem LC setup described above was used for the comparison of N2 and *daf-2* worms. Previous research on *daf-2* has shown that metabolites are from different metabolite classes that show significant differences. The branched-chain amino acids leucine, isoleucine, and valine are

the most prominent examples. However, also different fatty acids and lipids have been linked to the changed metabolism in *daf-2* mutants. Data from the used tandem-LC-MS setup separated wild type from mutant worms based on metabolites detected in the entire chromatographic range or only in the RP or HILIC parts.

5.2.1. Material and Methods

5.2.1.1. Instrument setup

The first prototype of this Tandem-LC system was realized on a Thermo Dionex Ultimate 3000 (Thermo Scientific, Dreieich, Germany) consisting of a Dual gradient pump with two vacuum degassers, a WPS sampler, a flow manager with two 10-port-2-position valves and a Variable Wavelength Detector. The instrument was coupled to a Bruker maXis UHR-ToF-MS (Bruker Daltonic, Bremen, Germany). The pump was able to deliver up to 3 solvents per pump. Connections between the individual modules were realized with stainless-steel Viper capillaries of 180 µm inner diameter (Thermo Scientific, Dreieich, Germany). A Kinetex C18 (2.1 x 30 mm, 2.6 µm, Phenomenex, Aschaffenburg, Germany) was employed as a trap column, while a ZIC-cHILIC (2.1 x 100 mm, 3 µm, Merck-Sequant, Darmstadt Germany) and a Kinetex C18 (2.1 x 100 mm, 2.6 µm, Phenomenex, Aschaffenburg, Germany), were employed as separation columns. Samples were stored at four °C in the autosampler, and injection was performed via full-loop injection of 20 µL. Columns were kept at 40°C. The exact flow path is shown in Figure 26. MS detection was carried out using the Bruker maXis UHR-TOF-MS equipped with an Apollo II ESI source in positive and negative ionization modes. Each run contained a segment of diluted ESI Low Concentration Tune Mix (Agilent Technologies, Waldbronn, Germany) for internal recalibration.

Separation on the RP part was carried out using water + 0.1% formic acid and ACN + 0.1% formic acid as eluents, while the HILIC separation was carried out with 95% H2O / 5% CAN + 5 mM ammonium formate and pure ACN. The exact gradient conditions are shown in Table 30.

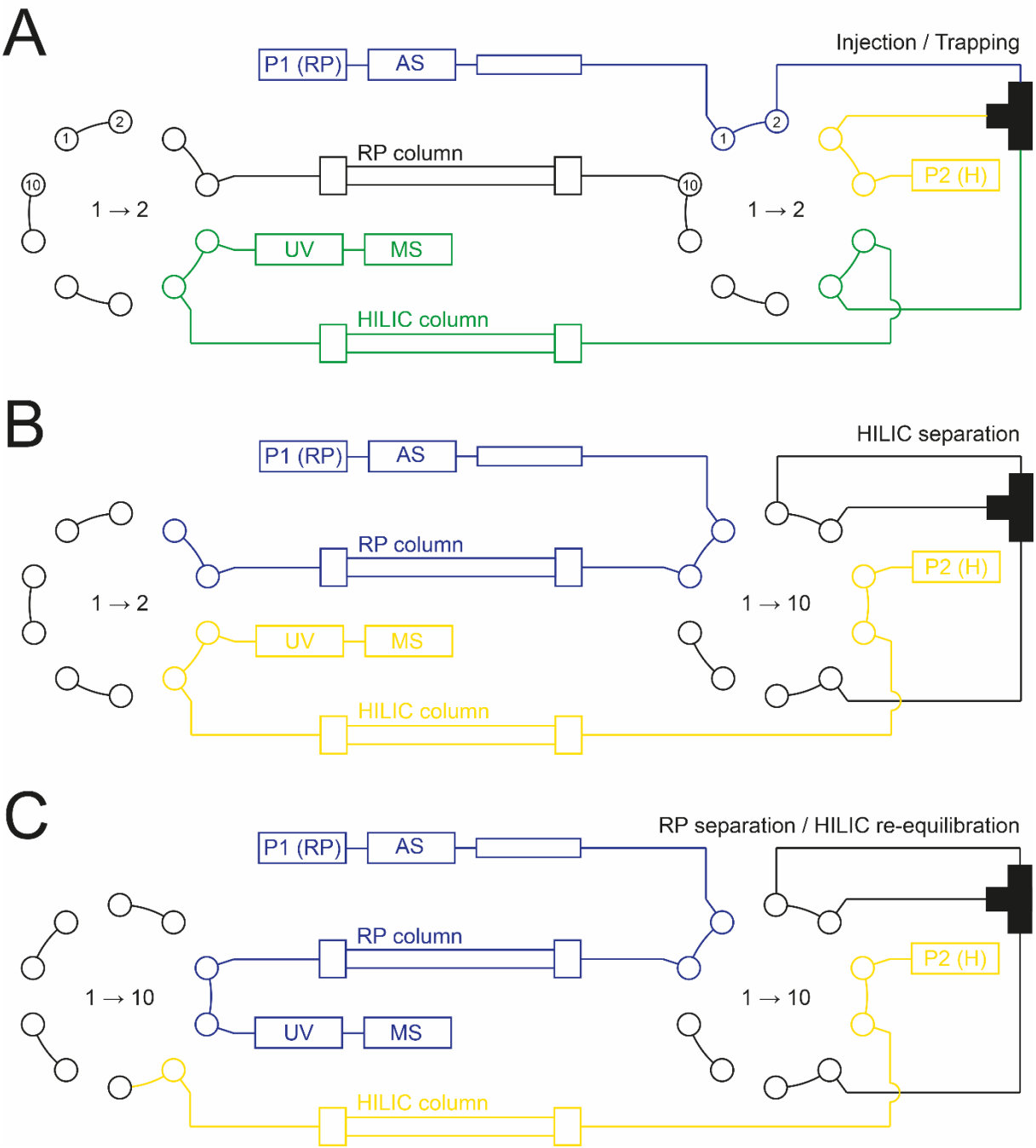


Figure 26: Schematics of the used tandemLC system. (A) During the injection and trapping phase the autosampler is connected to a short RP trapping column, the void volume is diluted with ACN and transferred to a HILIC column on which polar metabolites are trapped. (B) After the trapping the right valve is switched bringing the RP trap and separation column in line and the HILIC pump directly connected to the HILIC column. First the HILIC separation is performed. (C) After the HILIC separation is finished, the left column is switched and the RP separation is performed, while the HILIC column is reequilibrated.

Table 30: Gradient program used for the separation of metabolites and C. elegans extracts. Valves are referred as in Figure 26.

Range	Time	RP Pump		HILIC Pump		Valve positions
		%B	Flow	%B	Flow	Left | Right
Trapping	0.0	2.5	50	100.0	350	1_2 | 1_2
	5.0				350	1_2 | 1_2
HILIC	5.1		50	100	400	1_2 | 10.1
	18			40		
	24			40	400	
	24.9		400			1_2 | 10.1
RP	25.0					10_1 | 10_1
	26	2.5		90.0	200	
	40.0	100				
	45.0	100				
(Re)Equilibration	48.0	2.5				
	60	2.5	400	90.0	200	10_1 | 10_1

5.2.1.2. C. elegans culture and extraction

C. elegans strains N2 and *daf-2(e1370)* were cultivated according to standard protocols on NGM using E. coli OP50 as the sole food source at 20°C. Worms were synchronized by bleaching and grown until the young adult stage. Worms were collected by washing them off the plate with M9 and two additional washing steps to remove any remaining bacteria. Samples were snap-frozen in liquid N2 and stored at -80°C. Extraction was performed according to Witting *et al.* [185].

5.2.1.3. Tandem-LC-MS setup for C. elegans analysis

The setup was similar to the one described above. However, the ZIC-cHILIC column was replaced by an iHILIC-Fusion(P) (100 mm x 4.6 ID, 5.0 µm particle size, HILICON, Umea, Sweden). Solvents were similar to the initial setup.

5.2.1.4. Data preprocessing

Data from metabolites standards were manually extracted using Data Analysis 4.4 (Bruker Daltonics, Bremen, Germany). Data preprocessing of *C. elegans* was performed in Genedata Expressionist for MS 13.5 (Genedata, Basel, Switzerland), which included *m/z*-recalibration, noise subtraction, chromatographic alignment, peak picking, and isotope clustering. Data was exported as a .gda file for further statistical analysis in the Analyst module or as .xlsx for processing in Microsoft Excel. MS^2 spectra were exported as a .mgf file.

5.2.1.5. Statistical analysis

Data were imported into the Analyst module and first normalized using an intensity drift normalization, followed by a Probabilistic Quotient normalization (PQN) using the averaged QC samples as a reference. Only features detected in all QC samples with an RSD < 30% were further analyzed. To compare the wildtype N2 and *daf-2* mutant, a Welch test was used and features with a log2 foldchange <-1 or > 1 and a p-value < 0.05 were referred to as statistically significant between the groups.

5.2.1.6. Metabolite annotation

Metabolite annotation was performed using the workflow developed in chapter 3.6. An in-house spectral library, the Fiehn HILIC, and MassBank spectral libraries were used for annotation on the MS^2 level. Additional annotation on the MS^1 level was performed using a database of metabolites curated from literature, the WormJam model, ECMDB, HMDB, and SMID-DB. Furthermore, *in silico* annotation using Sirius 4.8.2 with CSI:FingerID and COSMIC [316]. Hits with a COSMIC score above 0.6 were considered high-confidence annotations. A molecular network was created using the Feature-Based Molecular Networking (FBMN) workflow on GNPS [300, 437]. Data were transformed to mimic the

tables exported from xcms 3 in order to be loaded into the FBMN workflow. The data were filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor *m/z*. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.02 Da, and the MS/MS fragment ion tolerance to 0.02 Da. A molecular network was then created where edges were filtered to have a cosine score above 0.6 and more than three matched peaks. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest-scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then searched against GNPS spectral libraries [299, 300]. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a score above 0.6 and at least three matched peaks. The DEREPLICATOR was used to annotate MS/MS spectra [474]. The molecular networks were visualized using Cytoscape software [475].

5.2.2. Results and Discussion

5.2.2.1. Reproducibility of Tandem-LC

Reproducibility in regard to the RT stability has been extensively validated by injecting twenty replicates of standards eluting in the HILIC or the RP part of the setup. Results are summarized in Table 31. The HILIC showed higher deviations in RTs compared to RP. However, this is typically also the case for isolated HILIC separations. Likewise, RTs were highly reproducible for different matrices such as plant extracts, wine, and *C. elegans* metabolite extracts. Detailed descriptions can be found in Hemmler *et al.* as well as in the master thesis of K. Wöhr [476].

5.2.2.2. Overview of C. elegans dataset

To validate if the developed TandemLC-MS setup is able to deliver valid metabolomics results, a comparison of N2 wildtype and *daf-2* mutant worms was performed. Samples were

analyzed in positive and negative ionization modes. The number of features after peak detection and isotope grouping, as well as after QC-based filtering, were as follows: In total, 13837 and 1308 features were detected in positive and negative ionization modes, respectively. After normalization and QC filtering, 8929 and 1092 remained for further statistical analysis. The lower number of features in the negative ionization mode can be potentially explained by the generally lower sensitivity in this mode. In positive mode, out of the 8929 features, 590 were detected in the trapping part, 5320 in the HILIC part, and 3019 in the RP part. In negative mode, out of 1092, 111, 762, and 219 were detected in the trapping, HILIC, or RP part, respectively.

PCA was able to separate both genotypes from each other in both ionization modes, and a Welch-test was used to identify statistically significant different metabolites. Metabolites with a log2 fold-change >1 or <-1 and a p-value < 0.05 were regarded as significantly different between the two genotypes. Nine hundred thirty-five were found to be higher in N2 and 385 in *daf-2* in positive ionization mode. Most of the significantly different metabolites were found in the HILIC part. From 385 specific for *daf-2*, 40 were detected in the trapping part, 315 in the HILIC part, and 30 in the RP part. N2-specific metabolites were 88, 621, and 226 in the trapping, HILIC and RP parts, respectively. In negative ionization mode, the proportions were similar. From 191 specific for *daf-2*, 13 were found in the trapping, 149 in the HILIC, and 29 in the RP part. From 91 metabolites specific to N2, the numbers were 13, 56, and 22.

5.2.2.3. Metabolite annotation and molecular networking

The obtained data is rich in MS^2 data since many different peaks were selected for fragmentation due to prolonged separation times and high efficient separation. In the negative ionization mode, 617 out of 1092, and in the positive ionization mode, 1859 out of 8929, features had a minimum of one MS^2 spectrum associated with it. The maximum number of MS^2 per feature was observed in negative mode, with 2636 for a single feature. In positive ionization mode, the highest number was 181. The lower coverage in positive mode can be explained by the high number of low-intensity features and higher density of features

in general. Annotations from different sources were combined into single consensus annotations. Additionally, GNPS FBMN was used. To correctly use this workflow, the output from Genedata Expressionist for MS 13.5 was used and converted to an output mimicking processing with xcms to be able to use the xcms-based workflow. R functions are used to directly read and convert .gda and .mgf files from the Expressionist output. The resulting feature table and annotated .mgf file were uploaded to the FBMN workflow.

Table 31: Reproducibility of RTs for standards of different polarity. Each standard was injected ten times.

Separation	Metabolites	RT (min)	RSD (%)
HILIC	Xanthosine	9.4	2.46
	L-Phenylalanine	10.6	2.14
	L-Tryptophan	11.1	1.87
	L-Valine	11.8	1.94
	L-Tyrosine	12.0	1.94
	L-Proline	12.2	1.95
	L-Asparagine	13.5	1.77
RP	C10 NAPS	30.2	0.07
	C11 NAPS	31.7	0.07
	C12 NAPS	32.7	0.00
	C13 NAPS	33.6	0.00
	C14 NAPS	34.5	0.13
	C15 NAPS	35.3	0.00
	C16 NAPS	36.1	0.06
	C17 NAPS	36.8	0.00
	C18 NAPS	37.6	0.00
	C19 NAPS	38.3	0.00

In negative ionization mode, 13 network clusters with more than five members were found, while in positive ionization mode, 57 were found. Since the network contains all edges, even across different parts of the tandemLC setup, the number of edges within one region or across has been counted (Table 32). Interestingly, the highest number of edges were observed within one range, e.g. RP to RP. However, no significant differences in the actual cosine similarity were observed. Similarities in FBMN are established based on spectral similarity using a modified cosine score which considers the difference between the precursor *m/z* and searches for fragments showing the same *m/z* difference. However, mass differences might not always be meaningful. Therefore, observed *m/z* differences were counted. In both ionization modes, most edges were found to be mass differences related to homologs series of lipid-like species such as 2.016 (H_2), 14.015 (CH_2), 28.031 (C_2H_4) or combinations of them (e.g. 26.015 C_2H_2). These might be mostly related to fatty acids or molecules having alkyl chains. Such molecules might be preferentially detected in the RP range. In negative ionization mode, 33 of the 53 edges related to CH_2 were found in the RP range, and 35 of 38 were related to H_2 in the RP range as well in negative ion mode. In positive ionization mode, 61 of 142 and 85 of 138 were found in the RP range for the same mass differences, respectively. Other mass differences with high counts were related to oxidation e,g. 15.99 (O). Results indicate that molecular families of similar molecules are eluting within a specific chromatographic segment.

5.2.2.4. Investigating known markers

First, the recovery of known markers previously reported to be different between N2 and *daf-2* were investigated. The most prominent example are the BCAAs leucine, isoleucine, and valine. All three could be identified on MSI level 2 with matches against library spectra from different libraries, and isoleucine was additionally annotated using CSI:FingerID with a COSMIC score of 0.56, slightly below the set threshold for automated analysis. Interestingly, none of them showed an increase in *daf-2,* as described previously. Another marker increased in *daf-2* worms is trehalose. Two peaks significantly upregulated in *daf-2* have been putatively assigned to be the $[M+H]^+$ and $[M+Na]^+$ adduct of trehalose, detected in the

HILIC part of the tandem LC setup. Betaine (or better glycine betaine) is a metabolite showing different trends between publications investigating metabolic changes in *daf-2*. In Martin *et al.,* betaine showed a substantial reduction upon *daf-2* mutation, while Fuchs *et al.* changes depended on the allele [74, 477]. Betaine showed lower levels in *daf-2* worms in the tandem LC-MS results. Likewise, glutamic acid has been shown to have lower levels in *daf-2* mutants. A highly significant change with a low p-value was found, but the fold change was only modest. GPC and O-PC were both downregulated in *daf-2* mutants in Fuchs *et al.*, while choline was up. Choline was down-regulated in Martin *et al.* Consistent with this, both were found to be downregulated in *daf-2* worms in this analysis as well.

Table 32: Distribution of edges in molecular network across the different separation ranges. Ranges were defined according to the time of the specific interval in the gradient time. Molecular networks were calculated using GNPS and the number of connections within a specific separation range or between ranges were counted. Generally, the most edges were found within a specific separation range, indicating that molecular families of similar molecules are eluting in the same segment.

From Range	To Range	negative mode	positive mode
HILIC	HILIC	283	873
	RP	30	64
	Trapping	10	51
RP	HILIC	11	29
	RP	258	1205
	Trapping	3	9
Trapping	HILIC	10	60
	RP	4	8
	Trapping	37	86

5.2.2.5. Changes in amino acids

In total, 18 amino acids were detected, 12 of them belonging to the proteinogenic amino acids. Six were significantly changed between N2 and *daf-2* worms. One of them is glutamic acid, as indicated above. Interestingly, the aromatic amino acids phenylalanine and tyrosine

were decreased in *daf-2* worms. In contrast to this result, Fuchs *et al.* found an increase, while no changes were observed by Martin *et al.*. Aspartic acid was in line with Fuchs *et al.* showing higher levels in N2 wild-type worms. Aspartic acid was not detected by Martin *et al.* Lastly, citrulline and pyroglutamic acid showed higher levels in N2 wild-type worms.

5.2.2.6. Metabolite features with highest changes

In the next step, features showing the highest changes between the two genotypes were investigated. One positive ionization mode example is ophthalmic acid, which is a tripeptide analog of glutathione in which the cysteine residue is replaced by 2-aminobutyric acid, found in higher levels in N2. Ophthalmic acid is discussed as a potential biomarker of oxidative stress. However, glutathione and glutathione disulfide showed no significant differences. Interestingly, aminobutyric acid is also increased in N2 worms. Further metabolites elevated in N2 were identified as Carn(3:0), Carn(5:0), and succinyl carnitine. All of them can be linked to the degradation of BCAAs. Though leucine, isoleucine, and valine were not found to be significantly different higher levels of these carnitines might indicate higher activity of BCAA degradation in N2 compared to *daf-2*. Carn(2:0) and Carn(4:0), in contrast, were not changed. Metabolites increased in daf-2 showed a high number of peptides containing proline.

In negative ionization mode, several fatty acids were detected and annotated via MS^1, formula calculation via Sirius, and checking of retention time trends. Out of the 11 detected, six were found in significantly higher levels in N2 worms (FA 18:1, FA 18:3, FA 18:4, FA 20:3, and two times FA 20:4) and one was only exclusively detected in N2 (FA(18:5)). Of these, FA 20:4 and FA 20:3 have also been found in higher levels in N2 by Castro *et al.* [121]. Similar results were also found by Lourenco *et al.* [75]. Higher levels of Trehalose-6-phosphate were found in *daf-2* worms, consistent with higher levels of Trehalose detected in positive ionization mode. Another molecule to be found higher in daf-2 was tyglu#2, a secondary metabolite from the SMID-DB. No reference spectrum has been deposited in known MS^2 databases, but a spectrum was available from the SMID-DB webpage. Manual

comparison of peaks showed an overlap of *m/z* 78.9591, 96.9696, and 223.0020. Spectral similarity was 0.8. More peaks were detected in the measured spectrum, which can help with interpretation. A fragmentation tree has been constructed using Sirius to help with identification. The two fragments at *m/z* 378.0955 and 360.0863 correspond to the loss of the 4-aminobenzoic acid moiety followed by the loss of the tyramine residue. Further fragments correspond to the sugar and phosphate group.

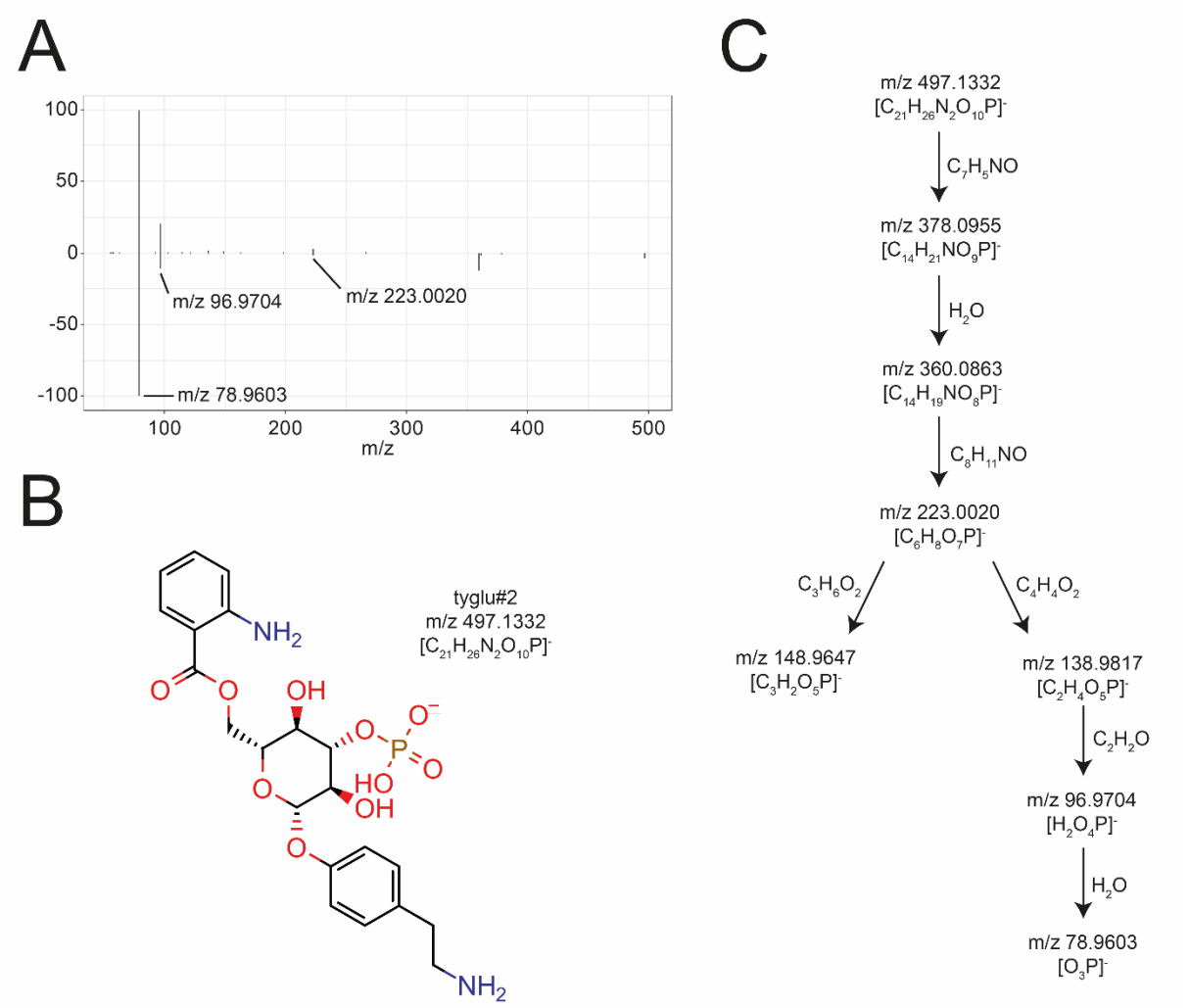


Figure 27: (A) Comparison of tyglu#2 reference spectrum from SMID-DB (upper panel) against the measured spectrum (lower panel). Three fragments matched, namely m/z 78, 96 and 223. (B) structure of tyglu#2. (C) Fragmentation tree computed by Sirius. The first two fragments, m/z 378 and 360, correspond to the loss of the 4-aminobenzoic acid moiety, while the next fragment also lost the tyramine residue. All further fragment belong to the sugar and phosphate structures.

5.2.2.7. Glycerophosphoethanolamides

One particular metabolite detected and down-regulated in daf-2 mutant worms was detected at *m/z* 470.23 at 31.28 min in negative ionization mode. A metabolite with the same *m/z* was

shown to be upregulated upon infection with *Pseudomonas aeruginosa* or *Salmonella enterica* [185]. Investigating the MS^2 spectra associated with this feature showed a pattern specific for glycerophosphoethanolamides (GPNEA), as described by Artyukhin *et al.* [478]. The fragmentation of this class of molecules has been described in this publication. The most abundant fragments identifying this class of molecules are a combination of *m/z* 78.9603, 96.9704, 152.9958, and 171.0063. GPNAEs are isomeric structures to LPEs. In contrast to GPNEAs, LPEs yield a fragment specific to the fatty acid attached in negative ionization mode. Interestingly, the molecule detected yielded lower intensities of *m/z* 96.9696 and different intensity ratios for the other fragments. In order to identify further members of this metabolite class, all MS^2 spectra were searched for spectra containing all of the following exact *m/z*: 78.959054 ($[PO_3]^-$), 96.969619 ($[H_2PO_4]^-$), 152.995834 ($[C_3H_6O_5P]^-$) and 171.006399 ($[C_3H_8O_6P]^-$). Using this, 79 additional features as potential candidates for GPNAEs were identified. Checking for the typical intensity patterns, finally, 23 remained. Further examination in Sirius revealed mostly the correct sum formulae, but often features were wrongly annotated as LPEs. In order to avoid the wrong annotation of the isomeric classes LPE and GPNAE, also the fragmentation of LPEs was examined. Based on the fragmentation, 16 GPNAEs could be identified. Positive ionization mode data was searched for the corresponding peaks, and 13 GPNAEs could also be identified in this mode as well. Similar to the negative ionization mode, LPEs, and GPNAE yield different fragments and can also be differentiated by their retention time. Using the logP as proxy GPNAEs are more polar, followed by LPE species. For example, the predicted logP of GPNAE 14:0 is 3.03, while for both LPE 14:0/0:0 and LPE 0:0/14:0 the logP is 3.65. This is also reflected in the RTs, as GPNAEs typically elute before the two LPE peaks. While LPEs typically show the loss of the PE headgroup in positive mode, GPNAEs mostly yield a fragment corresponding to the respective N-Acylethanolamide with an additional loss of H_2O. GPNAEs formed a dense cluster in the GNPS network with mass differences related to the elongation of the fatty acid tail. Of the 16 identified, seven were significantly different between N2 and *daf-2*. 4 were specific for *daf-2* and 3 specific for N2. These three all contained PUFAs. The feature

with *m/z* 470.23 contains an unusual PUFA FA 18:5. To verify if this fatty acid is possible, the presence of free FA18:5 was checked. Indeed, a low-intensity peak was detected, but since it was very low in abundance, no MS^2 with good quality was collected. However, the identity could be confirmed by using a trendline along the RT for all identified members of the C18 fatty acids. Besides the GPNAEs also, their glycosyl variants (HexGPNAEs) have been described. Described fragmentation includes the previously known fragments and additional fragments derived from the sugar moiety (59.01256 ($[C_2H_3O_2]^-$), 71.01260 ($[C_3H_3O_2]^-$), 78.95783 ($[PO_3]^-$), 89.02332 ($[C_3H_5O_3]^-$), 96.96854 ($[H_2PO_4]^-$), 101.02341 ($[C_4H_5O_3]^-$), 113.02357 ($[C_5H_5O_3]^-$), 119.03416 ($[C_4H_7O_4]^-$), 152.99548 ($[C_3H_6PO_5]^-$), 171.00616 ($[C_3H_8PO_6]^-$), 333.05972 ($[C_9H_{18}PO_{11}]^-$)). Interestingly, GPNAEs and HexGPNAEs were not connected within the GNPS network and formed separate clusters. The addition of the hexosyl moiety changed fragmentation in such a way that no similarity could be found.

GPNAEs are potential intermediates on the synthetic route to produce NAE. NAEs have been shown to mediate the effect of diet on lifespan, for example [101]. NAEs have been detected as [M+H]+ and [M+Na]+ adducts in positive ionization mode. Their fragmentation spectra could confirm several. Several of them have been found to be significantly higher in N2 worms. The highest fold change was found for NAE 20:5, the species also studied by Lucanic *et al.* [101].

Differences in the changes between GPNAEs and HexGPNAEs can be observed; for example, GPNAE 18:5 is significantly higher in N2, while HexGPNAE 18:5 is not significantly changed. Additionally, species differing in double bonds are often differentially regulated, e.g. GPNAE 16:2 and HexGPNAE 16:2 are higher in *daf-2,* and GPNAE 16:3 and HexGPNAE 16:3 are higher in N2. Consistent with this, NAE 16:3 and NAE 18:5 are also higher in N2. In the case of the 20:5 tail, no GPNAE nor HexGPNAE have been detected. Levels might be low since they might be directly metabolized to NAE 20:5, which seems to have an important signaling function.

No metabolic pathway has been for the biosynthesis of NAEs in *C. elegans*, except for a hypothetical one also added to WormJam. This pathway integrates multiple levels of regulation. The biosynthetic precursors for all metabolites are PEs and NAPEs. Fatty acyls are derived from PCs, which are transferred to the PE headgroup. Interestingly the different GPNAEs and HexGPNAE are based on unusual fatty acyls, such as 16:2, 16:3, and 18:5. Helf *et al.* also detected this species in a recent study [454]. Furthermore, different fatty acids and NAEs have been identified, suggesting that the entire network is more complex than expected. 1-glycero-3-phospho-hexose as a potential degradation product of HexGPNAE was also detected in the HILIC part of the tandem-LC setup but was not significantly different between the conditions.

5.2.3. Conclusion

Increased coverage of metabolomics and lipidomics methods represents an important topic to enable a truly comprehensive analysis of the metabolome and lipidome. Different ways, such as parallel LC, tandem LC or two-dimensional LC, have been used to achieve this. The presented optimized tandem LC setup combined RP and HILIC separation to enable the analysis of the polar and non-polar metabolome from a single injection. The method utilizes the column coupling to fractionate the metabolome and use the metabolites eluting in the void volume of an RP separation and transfers them to a HILIC separation. The system was optimized in such a way that the two separation methods could be performed independently from each other. This first allows for improving the instrumentation's duty cycle because the HILIC column is re-equilibrated after the separation while RP separation is performed.

Although the setup involves several switching valves and connections, it was proven during routine analysis to represent a robust alternative to classic separated separations. The suggested setup is easier to use in comparison to 2D-LC setups, for example. Typically for metabolomic or lipidomics analysis, multiple heart-cutting or comprehensive 2D-LC are used. In both cases, the second-dimension separation characteristics need to be optimized to scale with the first dimension. Multiple heart-cutting modules to park peaks in sample loops have

been developed, enabling the decoupling of the first and second dimensions. In the case of comprehensive 2D-LC the major limitation is the data analysis due to the high dimensionality (1D-RT, 2D-RT, *m/z,* and intensity). Therefore, advanced data processing methods, similar to ion mobility software, are required. In contrast, the data obtained from the tandem-LC setup is of lower dimensionality and can be processed similarly to standard 1D-LC.

Compared to individual separation on separate RP and HILIC methods, particular metabolites either elute in the trapping, RP, or HILIC part. No duplication of metabolites has been found so far, which makes metabolite annotation and data analysis easier. The use of *C. elegans* metabolite extracts shows that the tandem-LC setup is capable of coping with complex extracts and generates valid biological insights. The developed tandem-LC setup allowed the comprehensive analysis of the *C. elegans* metabolome from single injections per ionization mode. Together with the developed annotation workflow, it allowed the analysis of the N2 and *daf-2* metabolome and the comparison. Additionally, dedicated workflows such as FBMN in GPNS helped further annotate specific secondary metabolites. The results obtained are in line with previously published results. Combining the data, it was possible to identify members of a new class of metabolites only recently described in *C. elegans* and not present in metabolite databases at the moment. These molecules were defined as GPNAEs and their glycosyl variants, HexGPNAEs. In the GNPS network, both formed two dense clusters, which were separated. The chemical modification caused a difference in fragmentation based on which both clusters could not be connected by using the modified cosine score from GNPS. In future different approaches might be required to combine such clusters and make (bio)chemical sense of them. Amara *et al.* recently described different networking approaches in metabolomics and how they can be potentially combined [307].

Besides these new molecules, previously identified metabolites, which also have been associated with *daf-2* mutation, have been detected. This included several amino acids, fatty acids, trehalose, and trehalose-6-phosphate. Results were contrasted against different publications performing similar comparisons. Interestingly, results were different between the

measurements performed with the tandem LC setup and the literature but also between the different publications. The metabolome is very condition-dependent, and the nutritional status of worms has a considerable impact. In all cases, *E. coli* has been used as food bacteria, but potentially the exact density on the plates, as well as the nutritional content, might not be comparable, leading to different results—for example, Spanier *et al.* compared C. elegans grown in different laboratories using lipidomics [449]. Although several general trends were conserved, the exact fold changes between the different laboratories could not reproduce. So far, no good axenic culture of *C. elegans* is possible, which would enable more detailed studies of C. elegans metabolism using a defined medium. Nevertheless, the results have shown that the *daf-2* mutation causes a major shift in metabolism, which is potentially also (at least in part) for the longevity of this mutant.

TandemLC-MS shows several advantages but is currently not used in a routine setup. First, the usage of multiple pumps and method development and optimization represent a significant barrier for laboratories aiming to adopt this workflow. Additionally, several additional valves are required, which adds a potential source for leakage or blocking. Troubleshooting of such a system can become cumbersome. The currently used setup uses HPLC columns. The next step will be the transfer to UHPLC format columns with sub-2-µm particles for increased efficiency. In such a system, the correct use of connection capillaries becomes very important not to lose the advantage of efficient separation. In a further step, such a system can be further miniaturized to enable the analysis of a lower amount of *C. elegans* per sample.

5.3. Single *C. elegans* fatty acid analysis

Fatty acids are an important class of molecules in *C. elegans*. Many critical biological pathways converge toward fatty acid metabolism and signaling. For example monounsaturated fatty acids have been linked to lifespan extension [479]. Total fatty acid profiles are typically measured using GC-FID or -MS methods. For analysis with GC, derivatization of fatty acids is required. Volatile fatty acid methyl ester (FAMEs) can be prepared directly from esterified fatty acids, e.g. with methanolic BF_3 [480]. Separation of FAMEs on a GC yields a high-resolution separation that is also able to separate different isomeric fatty acids. However, if isotopic labeling data of fatty acids shall be analyzed the fragmentation of PUFAs with more than 18 carbons in the EI source yields many different peaks, which can be typically not deconvoluted into isotopic distributions [123].

Furthermore, the sensitivity of this method is limited. Typically, several thousand worms are used for this type of analysis [252]. CI allows the analysis of intact fatty acids and is generally used with high-resolution instruments such as GC-CI-ToF-MS [105]. Other possibilities include the study of free fatty acids. However, the isolation of free fatty acids requires more sample preparation and great care since fatty acids can be present in different chemicals and surfaces, which increases the background. Analysis of nonesterified fatty acids (NEFAs) can be performed using LC-MS/MS in negative ionization mode [481]. However, free fatty acids in negative mode show only very limited fragmentation, especially for saturated species. Therefore often, only pseudo transitions in MRM mode are used.

Derivatization can increase the sensitivity of the fatty acid analysis. For example, most MS show higher sensitivity in positive ionization mode, but fatty acids are only amenable to this mode after derivatization. Charge-reversal derivatization making fatty acid analysis in positive ionization mode is therefore of great interest. One possibility is the derivatization with 2-dimethylamino ethylamine (DMED), which has been widely used [482, 483]. Another example was developed by Bollinger *et al.* as a charge reversal derivatization using N-(4-aminomethyl phenyl)pyridinium (AMPP) for the analysis of eicosanoids [484]. This

derivatization reagent converts the negative carboxylic group into a permanently positively charged group, allowing high-sensitivity detection. This derivatization technique has also been used for the analysis of fatty acids using LC-MS/MS or direct infusion [485, 486]. A similar derivatization strategy, based on N-(4-1'-aminoethylphenyl)pyridinium (AEPP), has been used by Li *et al.* to study the physiological function of FATH-1 in *C. elegans* [487]. FATH-1 produces 2-hydroxy fatty acids used in sphingolipid biosynthesis.

Since charge-reversal derivatization yields increased sensitivity, it raises the possibility of analyzing fatty acids from a low amount of *C. elegans* or even single worms. Here the extraction, saponification, derivatization, and detection of fatty acids from single *C. elegans* were developed and tested. The final method allowed the detection of several fatty acids, including different isomers of odd-numbered iso-branched-chain fatty acids.

5.3.1. Material and Methods

5.3.1.1. Chemicals

HPLC-grade hexane, LC-MS-grade methanol (MeOH), acetonitrile (ACN), and formic acid were obtained from Sigma-Aldrich (Sigma-Aldrich, Taufkirchen, Germany). Water was purified using a Merck Millipore Integral water purification system with a resistance of 18 MΩ and TOC < 5 ppb. Hydrochloric acid and potassium hydroxide were of analytical grade and also obtained from Sigma-Aldrich.

5.3.1.2. C. elegans culture

C. elegans were cultured according to standard protocols using NGM and *E. coli* OP50 as the sole food source. Worms were grown to the adult stage and then individually picked to an empty NGM plate to allow them to empty their guts. After a few minutes, worms were directly picked into the extraction solvent, as described below.

5.3.1.3. Total Fatty acid extraction

Potential contaminating fatty acids were washed off all glass material with LC-MS-grade MeOH. Adult worms were directly picked into 50 µL 1M KOH in MeOH which was placed in a total recovery LC-MS vial (Waters, Eschborn, Germany). To avoid contamination, the worm

picker was first heated in an ethanol flame and then washed in LC-MS grade MeOH before individual *C. elegans* were picked. Samples were closed tightly, and fatty acids were extracted at 60°C for 1 hour. After cooling, 1 µL of concentrated HCl was added to neutralize fatty acids, and 50 µL of H_2O was added. Next, fatty acids were extracted into 500 µL of hexane by shaking. After phase separation, the hexane was transferred to a fresh glass vial, and the solvent evaporated to dryness.

5.3.1.4. Fatty acid derivatization

Extracted fatty acids were derivatized with the AMP+ kit (Caymen Biochemicals, Michigan, USA). Instead of the original procedure, only half of the indicated volumes were used. To the dried fatty acids, 10 µL ACN/DMF, 10 µL EDC, 5 µL HOBt, and 15 µL AMP+ derivatization reagent were added. The sample vials were sealed tightly, and the reaction was conducted at 60°C for 30 minutes. After cooling, the vial was directly used for injection

5.3.1.5. UPLC-MS analysis

Derivatized fatty acids were separated on a Waters Acquity UPLC (Waters, Eschborn, Germany) using a Waters Acquity UPLC BEH C8 column (150 mm x 2.1 mm ID, 1.7 µm particle size) (Waters, Eschborn, Germany) and a 100% H_2O + 0.1% formic acid and 100% ACN + 0.1% formic acid as eluents. The column temperature was set to 40°C and flow rate to 0.25 mL/min. The following gradient was used: 60/40 at 0 min, 60/40 at 2 min, 0.1/99.9 at 20 min, 0.1/99.9 at 25 min, and 60/40 at 25.1 min. Detection was carried out using a Bruker maXis UHR-ToF-MS (Bruker Daltonic, Bremen, Germany) in positive ionization mode.

5.3.2. Results and Discussion

5.3.2.1. Method development and validation

The commercially available AMP+ kit from Cayman Chemicals has been selected for charge-reversal derivatization. It uses AMPP as a derivatization reagent and supplies all necessary chemicals. AMPP showed high gains in sensitivity in previous applications [484, 485]. In the first step, fatty acids are activated using 1-Ethyl-3-(3-dimethyl aminopropyl)carbodiimide (EDC) and 1-hydroxy benzotriazole to form an active ester. This active ester is then

converted into a stable amide with AMP (see Figure 28). A significant advantage of this method is a relative increase in mass and hydrophobicity to a certain extent that also allows the measurement of short-chain fatty acids such as butyrate [488]. First, free fatty acids were used to test the derivatization efficiency. As suggested, reactions are complete with over 95% yields (data not shown).

Chromatographic separation of derivatized fatty acids was optimized by testing different reversed-phase columns. The preselection of columns indicated that a BEH C8 worked best and could cover the complete anticipated fatty acid range from FA 12:0 to FA 30:0. Two different gradients were tested. An essential aspect for the analysis of *C. elegans* fatty acids is the separation of straight- and branched-chain fatty acids. Therefore, different free fatty acids were derivatized in three mixes, and separation was tested. The first gradient was a linear increase from 40% to 99.9% eluent B, while the second gradient first increased from 40% to 60%, followed by a second step from 60% to 99.9%. Already the first gradient was able to separate FA 15:0 and FA 14:0(13Me) as an example of isomeric species. However, the second gradient increased the resolution of isomeric species and was therefore selected for all further investigations.

First, the elution behavior of fatty acid standards was studied. The first part of the gradient roughly covers the range until C18:0. In this range, an increase in carbon chain by C_2H_4 yields a retention time difference of about 3.7 minutes. Fatty acids with a higher number of carbons elute in the second, steeper part and only have a difference of 0.2 minutes. One additional double bond reduced the retention time by about 2.3 minutes for the first and the second double bond, while the effect from the third onwards was dependent on the position of the double bond. For example, the series of FA 18:0, FA 18:1(9Z), FA 18:2(9Z,12Z), FA 18:3 (9Z,12Z,15Z), and FA 18:3 (6Z,9Z,12Z) eluted at 18.5, 16.2, 13.9, 11.9 and 12.4 minutes respectively.

Next, fragmentation spectra of the fatty acid derivates were examined. In the original article by Wang *et al.*, extensive fragmentation was described [486]. However, this analysis

observed no fragmentation, potentially because 35 eV is too low to yield the common fragments described. Instead, very low-intensity peaks at These fragments allow identifying peaks as derivatized fatty acids and, together with RT trends, different species for which no reference standard was measured can be identified.

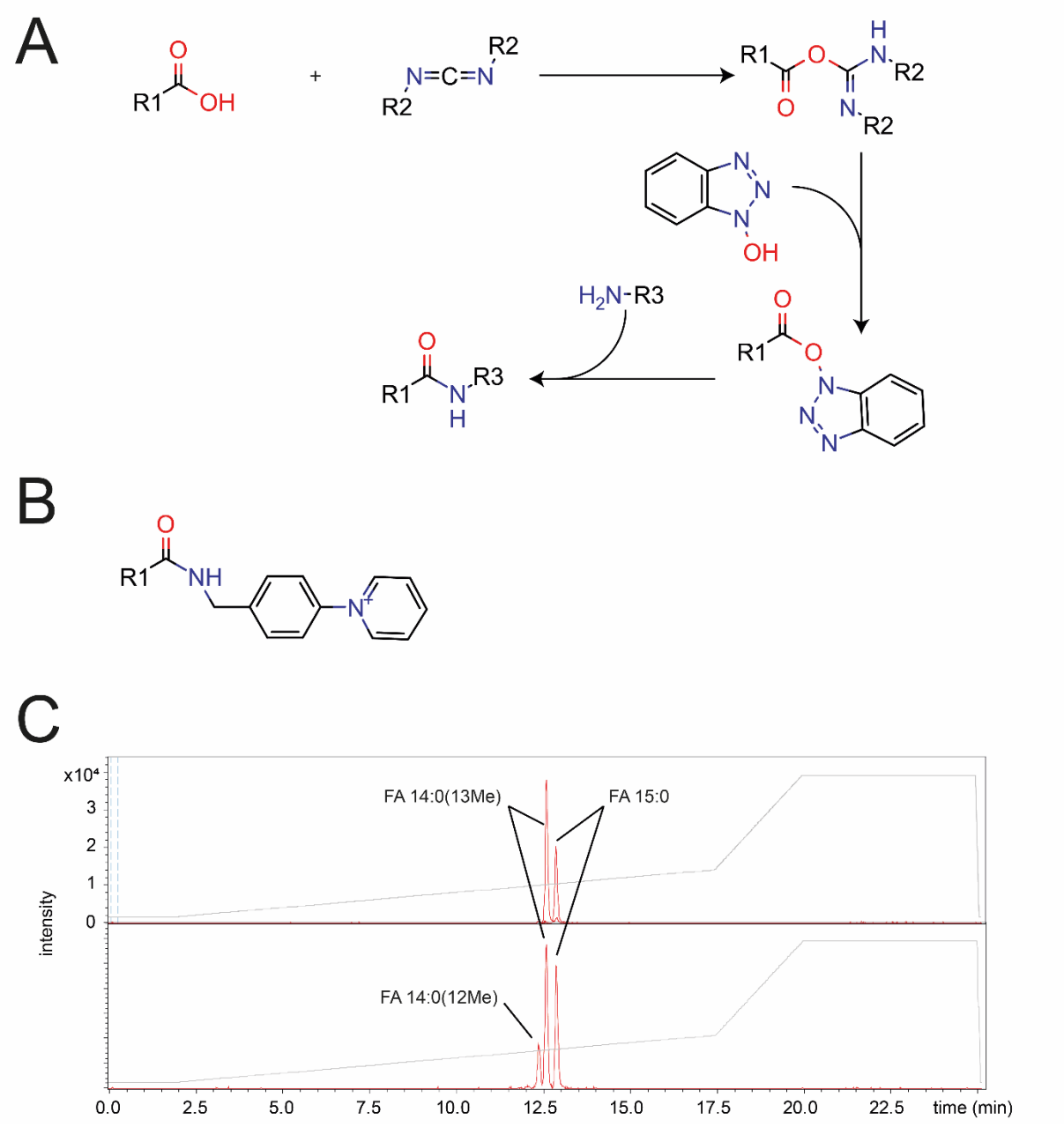


Figure 28: (A) Scheme of derivatization reaction used in the AMP+ kit. An active ester is formed, which then reacts with amino groups to form an amide. (B) Reaction product of AMP+ kit. The permanent cation allows sensitive detection in positive ionization mode. (C) Extracted ion chromatograms related to FA 15:0 and isomers. Good separation of FA 14:0(13Me) and FA 15:0 was achieved for the reference standards (upper panel) and in C. elegans samples (lower panel). Additionally, FA 14:0(12Me) was detected in C. elegans samples.

In the following step, the derivatization conditions for single *C. elegans* were tested. A single one-step extraction and saponification using methanolic KOH were selected to avoid

potential losses of material due to multiple steps and excessive handling. After neutralization, fatty acids were extracted into hexane and dried. Only half of the volumes originally described in the AMP+ kit were used not to dilute the sample too much. This procedure allowed still handling of low volumes with standard pipettes. After derivatization and cooling, the sample was directly injected into the UPLC-MS system. Results indicate that this procedure yields detectable signals from different fatty acids and was applied to multiple worms, as described below.

5.3.2.2. Application to single C. elegans and heterogeneity in fatty acid levels

As a proof-of-concept, the final method was used to profile fatty acids from several individual worms. Young adult worms from a mixed-stage culture were first picked to an empty NGM plate with no bacteria to remove excessive bacteria and defecate bacteria in the gut. Worms were then picked individually into the extraction solution and treated the same way, for this first proof-of-concept, in total, ten worms were processed.

Theoretical *m/z* values for *in-silico* derivatized fatty acids have been used to search for the corresponding peaks. In total, 15 different fatty acids were identified. Peak areas were used and normalized to the sum of all peak areas to normalize potential differences during sample preparation and measurements. The primary fatty acid detected was palmitic acid (FA 16:0), mostly over 50% of the total peak area, followed by stearic acid (FA 18:0). Interestingly, three different isomeric peaks for FA15:0 and FA 17:0 were detected, potentially representing anteiso, iso, and straight chain variants. Furthermore, two isomers of FA 18:1 were detected. However, it is important to mention that no absolute quantification has been performed. Therefore, the numbers are only rough indicators. Though derivatization might help normalize ionization behavior, differences in the eluent composition and the fatty acid itself can lead to differences. Therefore, a comparison between fatty acids is not made, and worms are only compared within one fatty acid.

Investigating the individual profiles, differences for all worms could be observed. To account for differences in the abundance of the individual fatty acids, the relative difference to the

average of all worms was depicted. Interestingly, Worm 1 showed large differences in the second peak of FA 15:0, the second peak of FA 17:0, as well as FA 16:1, the first peak of FA 18:1, FA 18:2, and FA 20:5. However, the latter was only detected in two samples. The peaks for FA 15:0 and FA 17:0 are based on the elution pattern of the iso-forms. It makes sense that both are increased since FA 17:0 is the elongation product of FA 15:0. Furthermore, the higher abundance of FA 18:1 and FA 18:2 and 16:1 are related as they can also be produced from each other. Similar increases can be seen in Worm 7, but not as extensive. Though worms generally tend to have similar fatty acid composition, selected worms show distinct profiles. Since these different profiles make sense biochemically, they represent real differences rather than extraction, derivatization, or measurement artifacts. How this difference would have translated, e.g., into differences in lifespan, remains elusive. Since the extraction for metabolomics and lipidomics is destructive, no further information on future lifespan could be obtained, nor can the worm be sampled a second time.

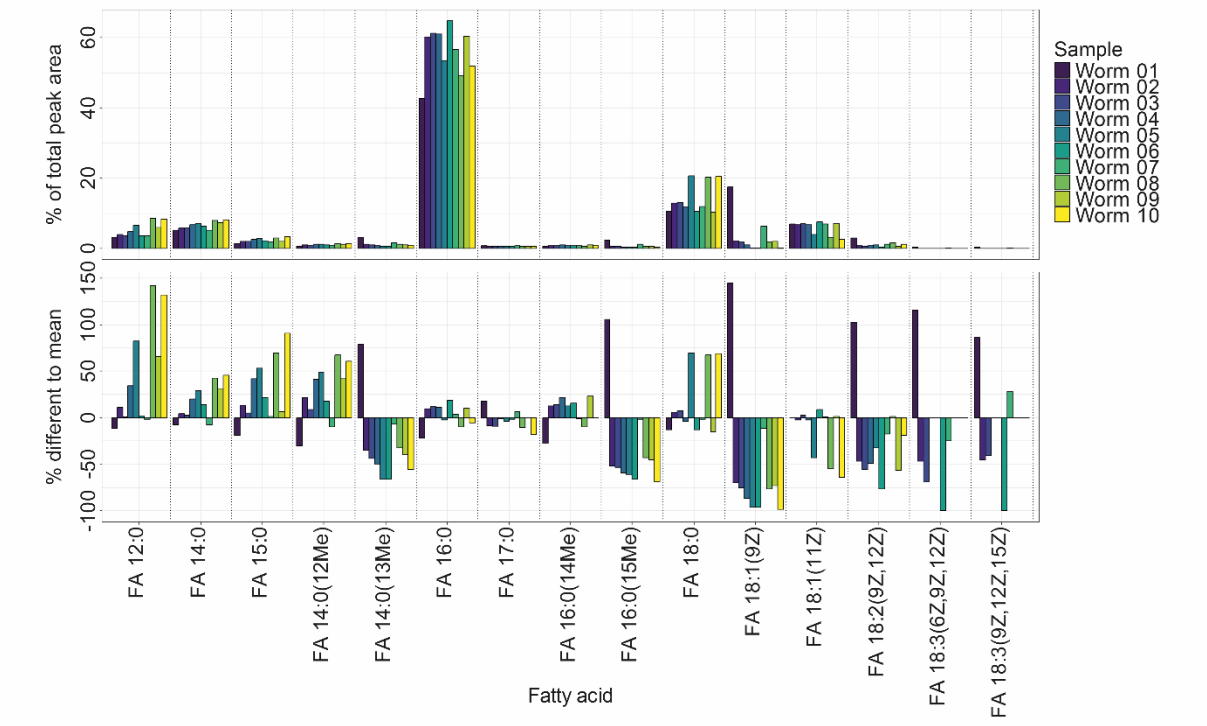


Figure 29: Fatty acid profiles from 10 individual worms obtained by LC-MS/MS analysis. The upper panel shows the % of total peak area for each fatty acids. Using the % of total peak area differences based on sample preparation from the individual worms are normalized. The lower panel indicates the differences of the single worms to the mean of all 10 worms. Worm 01 shows particular differences in branched chain fatty acids as well as unsaturated species.

5.3.3. Conclusion

Fatty acids are typically analyzed using GC-FID or GC-MS but can also be analyzed with LC-MS. However, native fatty acids analyzed in negative ionization mode show none to only weak fragmentation. Therefore, charge switch derivatization was developed to enable a more sensitive analysis of fatty acids in positive ionization mode. This derivatization adds a constant group, yielding a constant fragmentation, which can be used for quantification. Additionally, the sensitivity is greatly enhanced.

This derivatization was applied to samples of single *C. elegans*. Total fatty acids have been extracted by boiling in methanolic KOH, and saponified fatty acids have been derivatized. The sensitivity allowed obtaining fatty acid profiles with 15 detected fatty acids. Chromatographic separation separated anteiso-, iso- and straight-chain fatty acids. Using ten worms as proof-of-concepts, their profiles were compared. Using the averaged profile of all as a reference, particularly one worm was found to have a different profile.

However, the number of samples is still quite low, so further experimentation will be required. A significant point for improvement is the switch to absolute quantification. This would allow the comparison between different fatty acids and the calculation of ratios between direct substrates and products, e.g., of elongation or desaturation reactions. In order to further increase the sensitivity of a column with a smaller inner diameter would be possible. While the development of such technology holds great promise to explore more interesting aspects of *C. elegans* biology, several questions remain. First, the heterogeneity of worms cannot be neglected at this level of analysis. So while the amount of biomaterial might be reduced to a single worm, the number of replicates has to be increased to obtain stable read-outs for statistical analysis. It remains unknown at the moment how many replicates might be required. The question is, what amount is sufficient and practicably feasible? Second, limited biological starting material also limits the number of possible analyses. In a worst case, it might be possible to analyze a single sample only once. This demands that the information obtained from this sample must be as comprehensive as possible. Lastly, due to the

destructive nature of the extraction, no further analysis of the lifespan of C. elegans can be performed. One possibility is to grow a parallel cohort of worms or use worms grown on the same plate as references. However, these cohorts must be linked. Potential read-outs are either different reporter genes or autofluorescence, for example.

- This page is left intentionally blank -

6. Conclusion and Outlook

6.1. *C. elegans* as model organism for metabolism research

C. elegans is one of the premier models to study metabolism and its effects on different phenotypes. Despite the fact that it is studied since a few decades now, surprisingly little is known about the complete metabolism and its regulation in health and disease. Metabolomics and lipidomic aim to close this gap by generating a comprehensive overview on metabolism from a given sample. By combining them with genetics and other functional assays new avenues for metabolism research are being created. While metabolomics and lipidomics are well-developed in human research with the availability of rich resources such as the HMDB such resources are missing for *C. elegans* [86, 87, 198, 489].

This work aimed to close this gap to enable the further development of *C. elegans* as model organism for metabolism research. This included the creation of a consensus GSMN, WormJam, for unified in silico analysis of the nematode's metabolism. Different GMSNs have been generated for *C. elegans*, but common to all they lack the *C. elegans*-specificity (e.g. the use of C17iso sphingoid bases or inclusion of ascaroside metabolism).

Still several major gaps have been identified. One particular example from this work is the biosynthesis of N-acyl-ethanolamides (NAE) in C. elegans. In the literature these metabolites have been described several times, e.g. in Lucanic et al. [101]. GPNAEs as intermediates of the biosynthesis NAEs have been detected was well [140, 478]. Furthermore, hexosylvariants of the GPNAEs, HexGPNAEs, have been described and detected in this work. However, the first steps of the biosynthesis remain to be elusive. Although tow N-acyl-phosphatidylethanolamine-specific phospholipase D orthologs, named nape-1 and nape-2, have been described, their activity has been only shown in vitro. However, the first step such as the N-acyl-transferase are missing. Especially, NAPEs have not been detected so far in C. elegans lipid extracts and due to their low levels in other organisms (typically less than 0.1% of total phospholipids) they might require additional sample preparation steps [490, 491]. Elucidation of the complete biosynthetic pathway might represent an important task given that several pathways converge on NAE signaling.

The combination of the WormJam in silico model, prediction of lipid metabolic pathways and curation of metabolites and lipids allowed to construct a first version of a C. elegans metabolome/lipidome database. As anticipated the metabolism of the nematode forms a dense network with several thousand possible metabolites and lipids. This database will server as blue-print for further investigations into the metabolism of C. elegans. Since for metabolites present in WormJam direct association with pathways are available, systematic analysis of metabolism will become possible.

While this represents a compound database, reference spectra for several C. elegans-specific metabolites are still missing in publicly available databases such as MassBank. Though several spectra a available via the SMID-DB database, they have to copied manually and data files for spectra matching need to be generated by the user. In case of lipid spectra this missingness can be circumvented by the prediction of spectra using rules from the fragmentation of known lipids. This has been proven previously to very successful for LipidBlast and the generated lipid spectra within this work showed good matching with known lipids and helped to identify several lipid species in C. elegans. Rule based annotation of lipids is also generally increasing in commercial software, e.g. Bruker MetaboScape, or in open-source software, e.g. mzMine 3.0.

Several possible next steps for the improvement of the WormJam GSMN are possible. First, it misses several compartments, e.g. the peroxisome. Refinement of reactions and their subcellular location will be an important next step. Since, C. elegans has differentiated tissues it will be important to incorporate this information as well. A first step in this direction has been performed by Yilmaz *et al.* [492]. Single-cell RNA sequencing has been used to reconstruct tissue specific metabolic pathways for seven different tissues. Models and flux analysis are consistent with previous knowledge on the metabolic function of these tissues. Lastly, interaction of C. elegans with its natural microbiome is an important and upcoming topic. Dirksen *et al.* presented CeMbio, a resource containing different culturable bacterial strains of the C. elegans microbiome [493].

6.2. Systems biology approaches in *C. elegans* metabolism

Experimental setups in C. elegans are becoming larger, with multiple conditions, genotypes, etc. While large-scale experiments such as genome-wide screens are typically combined with other read-outs such as fluorescence microscopy. In future, more and more applications of C. elegans in systems biology approaches of metabolism will be observed. Genomics approaches are currently capable of using low number of worms cultivated in for example 96-well plates. Metabolomics and lipidomics offers a multiparametric read-out, which in future can be combined with large-scale experiments. While the power of MS analysis is increasing, several obstacles need to be removed. Throughput is the major factor which needs to be optimized. Several examples for successful application exist.

For example for the analysis of a *E. coli* KO library a high-throughput FIA-ToF-MS has been developed [48]. Throughput is an important topic in metabolomics and lipidomics and methods are constantly improved in terms of speed [494-496]. Run times below 5 minutes are now feasible. Such approaches are required to embed C. elegans metabolomics and lipidomics in the systems biology landscape to be on par with other approaches such as imaging or phenotypic screens. Calculating a library of 15.000 theoretical knockouts (not all KOs are viable or develop normal), a method using RP column with 15 minutes per ionization mode would require 156.25 days pure measurement time per ionization mode. No included are additional required system samples, such as blanks, QC or reference samples. Reduction to 5 minutes per ion mode yield 104 days measurement time, and a 3-minute method 63 days. With such a timescale the screening of multiple conditions becomes possible.

An important factor is sample preparation as well as QC aspects. Within each batch several controls that enable comparability between the different batches are required. Furthermore, reference materials shall be used for control of each batch. A recently introduced reference material for C. elegans can help to improve data quality and comparability [462].

However, interpretation of data becomes a bottleneck. Using the approaches developed in this work for metabolite annotation can significantly speed up the process of data analysis. Furthermore, the workflow consists of defined steps and the is reproducible. Together with version-controlled libraries and database it is allows the consistent annotation and identification. The annotation of novel, unknown metabolites and lipids will remain a major bottle. Since, *C. elegans* can be grown in larger quantities the purification of metabolites and lipids for NMR is typically possible using analytical or preparative scale chromatography. Pathway level interpretation of obtained data will become an important topic. GSMNs such as WormJam, will allow the large-scale interpretation, but no tools are available for example for pathway enrichment in *C. elegans*. MetaboAnalyst for example supports pathway enrichment for *C. elegans*, but the origin of the background pathways is unclear [497]. Since the entire metabolome and lipidome is currently not known it is hard construct a correct background dataset for enrichments.

6.3. *C. elegans* metabolome and lipidome database drafts

Reference databases for species-specific metabolomes and lipidomes are becoming more popular and available. The species-specificity circumvents over-annotation of metabolomics data using generic large databases. Different trends are visible, either single databases dedicated to a single organism are created or larger databases adopt some kind of information which can link a metabolite or lipid to an organism. One recent example is LOTUS, which stores information of natural products and their links to producing organisms [498]. Other examples such as SMID-DB for C. elegans specific secondary metabolites have been mentioned earlier.

The chemical curation of the WormJam GSMNs as well as the lipid prediction workflow allowed to generate a first draft of the *C. elegans* metabolome and lipidome database. Comparison against detected metabolites has shown that still a large part is missing, which is especially true for secondary metabolites. This represents an interesting field and new secondary metabolites are described on a routine basis, e.g. by the Schroeder group.

6.4. New analytical approaches

The increased demand in metabolomics and lipidomics analysis will require the refinement of methods such as the ones presented here as well as the development of new methods. In parallel to the increase of number of samples a reduction in the biological start material (due to different culture conditions in 96-well plates for example) will be required. Together with the improvement of analytical capabilities this will create new avenues in *C. elegans* metabolomics and lipidomics. A major point is the increased sensitivity required for the analysis of smaller amounts of biological material. Miniaturization, especially of the LC component is a major target point for improvements. While nanoLC is readily used in proteomics and represents the standard technique, its use in metabolomics and lipidomics is underexplored. Ultimately this will allow to determine the metabolome and lipidome of a single worm, similar to the determination of a single worm proteome [188].

In this work a tandem-LC setup was developed which combines RP and HILIC separation in a single setup. This allowed to increase the coverage of metabolites from a single injection. This setup has been successfully applied to study metabolic differences between N2 and daf-2 worms. This setup has been realized on an older model of the LC system allowing a maximum pressure of only 400 bars. A prototypic setup with higher pressure ranges has been tested in collaboration with Thermo Fisher Scientific. This setup was working and first results have been obtained with *C. elegans*. However, the injection volumes are still rather large and requires larger amounts of *C. elegans* and therefore make it only partially useful for large-scale studies such as genetic screens. Ultimately realizing this tandem-LC setup in form of a nanoLC-MS system will yield sensitivity and coverage for the analysis of single worms.

Generally the switch towards smaller inner diameters increases the sensitivity of analysis (compare Table 33). For example moving from a 2.1 mm ID column to a nanoLC column with 75µm ID and theoretical sensitivity increase of 743-fold can be achieved. With MicroLC-systems an increase of 4.2 to 41 fold is possible, but stability in much higher compared to

nanoLC. MicroLC systems have been used in the past for example for the analysis of polar and non-polar metabolites and lipids [494, 499-501]. Even nanoLC has been used for the analysis of lipids [502, 503]. However, nanoLC-MS suffers from lower (chromatographic) reproducibility compared to analytical flow LC-MS. In proteomics as shifts towards the use of microLC-MS is visible since modern MS systems offer higher sensitivities on the MS side. The lower flow rates also aid the ionization process in ESI source potential yielding a further gain in sensitivity. However, exact increases in sensitivity are specific to metabolites and metabolite classes and can be very different among then. Potentially, metabolomics, lipidomics and proteomics will converge towards the use of micro-flow systems with increased robustness. Especially robustness is required for large-scale analysis such as genome-wide screens.

Table 33: Theoretical comparison of different columns dimensions. Columns with 2.1 mm have been set as standard and the table is scaled to this columns.

ID (mm)*	Flow Rates	Injection Volume	Sensitivity	Mobile Phase Consumption	Typical Flow Rates
4.6	4.8	4.8	0.2	4.8	>= 1 mL/min
3	2	2	0.49	2	0.4-0.6 mL/min
2.1	1	1	1	1	0.2-0.4 mL/min
1	0.22	0.22	4.2	0.22	50-100 µL/min
0.5	0.056	0.056	17	0.056	10-20 µL/min
0.32	0.022	0.022	41	0.022	5-10 µL/min
0.18	0.0078	0.0078	129	0.0078	1-5 µL/min
0.075	0.00136	0.00136	743	0.00136	0.2-1 µL/min

For example, in Häussler *et al.*, which data has been used in here a genome-wide RNAi screen has been performed. In this screen imaging has been performed and 10-20 L4 larvae of the F1 generation were used. Lipid profiling was only performed after hit-verification using larger cultures and populations. Direct integration of lipid analysis in the screen would have

been of great interest. First, potential candidates, which might not influence the used reporter, but lipids might have been identified and lead to a larger number of potential hits for further downstream experimentation. Lipidomics in this regard can be seen as multi-reporter read-out. Instead of focusing on a single GFP-tagged protein or a selection of fluorescently labeled proteins these multiple read-outs might lead to a better understanding and a more systems biology (holistic) view. Second, differences between culture conditions, which are known to influence the outcome of lipid profiling experiments, can be reduced [449].

6.5. Further analytical requirements

Although many new metabolites and lipids could be identified in this work and large-scale dereplication of know metabolites has been performed, still many unknown metabolites remain. One particular reason is that not all features have an associated MS2 spectra, especially lower intensity features or low quality. Though newer generation of MS instruments have increased duty cycles and improved ion optics. The way of data acquisition has not fundamentally changed within the last years. DDA is still the dominant way to collect MS2 spectra for detected m/z values. However, in many cases no further information is used to select precursor except for abundance and potential isotopic pattern. Major problems are the detection of adducts or potential in-source fragments as well as the selection of the optimal fragmentation energy. Especially the later is important for untargeted analysis, since the structures are not known, the optimal collision energy is often also not known. Furthermore, in silico tools such as Sirius and CSI:FingerID perform better when information from multiple collision energies is supplied. Therefore, either multiple spectra from the same precursor with different collision energies (10, 20 and 40 eV for example) or ramped or stepped spectra are collected. The first one has the advantage that induvial spectra are collected and can be also used for comparison against library spectra, which are often collected at individual collision energies. However, the number of precursors covered is reduced to increased time required for the multiple spectra. When using a low amount of material potentially not enough material is available for reanalysis.

DIA might represent a promising alternative approach to DDA in metabolomics and lipidomics in future. First results presented here show that in combination with ion mobility it is a useful tool for identification of lipids. However, the presented study focused on specific lipid class. The used DIA approach represents a single window approach combined with ion mobility. Other possibilities include multiple sequential windows or a scanning approach. Though the potential applications are broad, DIA is not widespread in metabolomics and lipidomics. The major reason is the limited data analytical possibility. Reconstruction of spectra represents a huge deconvolution problem, which is similar to EI-GC-MS. However, the better chromatographic separation in GC allows easier reconstruction. Most approaches so far are limited to targeted analysis of previously known and characterized substances. One possibility is the DIAMetAlyzer approach [504]. Results have been shown to be similar to manual processing. In case of unknown metabolites for which no reference spectrum is available reconstruction of fragmentation spectra from raw data becomes complicated. Reconstructed spectra will be different to DDA spectra. Dependent on the used window size it will contain isotopes as well as represent a chimeric spectrum of different adducts, if they fall into the same window or bin. This complicates analysis and requires further deconvolution approaches. Newer DIA approaches such as SONAR or scanning SWATH might represent suitable alternatives, since they are using small windows of only a few Daltons. In proteomics the timsTOF is often used, which allows combination of ion mobility and DIA. Several approaches such as diaPASEF, Synchro-PASEF and midiaPASEF have been developed to deconvolute complex peptide spectra [505-507]. However, none of these approaches has been used for the analysis of metabolites or lipids so far. Generally, DIA is a useful approach to create a “digital archive” of a sample and data can be retrospectively analyzed once new hypothesis have been generated.

In addition so far only metabolites detected under the employed conditions can be analyzed. ESI is employed in the majority of metabolomics and lipidomics analysis. However, certain metabolites might not be ionized or only show low intensity in ESI. APCI for example is a suitable alternative for the analysis of steroids and other compound classes [508, 509].

Structure elucidation and dereplication of small molecules remains an important question in the near future. Fast dereplication, meaning the identification of previously detected and quantified metabolites and lipids, will be the main task to allow to focus on truly novel structures to be elucidated. In proteomics for example Parallal Database Search Engine in Real Time (PASER) is used for real time identification of peptides during an analytical run. Similar approaches will be required in metabolomics and lipidomics. Furthermore, several new approaches are required to speed up the identification of novel metabolites. This requires more analytical depth, such as novel fragmentation mechanisms. Electron activated dissociation (EAD) has been recently described for the analysis of lipids [510]. This type of fragmentation allows the determination of the position of fatty acyls as well as the location of double bonds reaching a deep lipid annotation. First results show that EAD might be also helpful for the annotation and identification of metabolites, e.g. identifying positions of hydroxy or methyl groups. Since this type of fragmentation produces radical ions, which are normally not observed under CID conditions, most of the software tools cannot be directly used. Similar to CID database need to be compiled to be used a training and test data for the development of new data analysis approaches. So far EAD has not been applied to C. elegans or C. elegans specific metabolites, but it can help especially with the analysis of secondary metabolites such as ascarosides.

GSMNs are striving towards the analysis of metabolism of context of differentiated tissues. However, for proving of predicted metabolic functions spatial approaches are required. So far C. elegans metabolomics and proteomics has been performed on the level of entire cultures with several hundreds to thousands of worms. Even when developments will allow the reduction towards a few tens of worms or a single worm, after extraction the information on tissue specificity is no longer available. MS imaging has been developed as tools to analyze metabolomics in spatial context. In combination with the ever-increasing sensitivity of MS instrumentation the spatial resolution for imaging experiments in increasing as well. Newest generation instruments of MALDI-MS reach now spatial resolutions of 5 µm, research instruments even lower. An adult *C. elegans* has a diameter of 65 µm. High spatial

resolutions will allow to analyze metabolites in context of specific organs. Lastly, single-cell proteomics and metabolomics have gained popularity. Using techniques such as fluorescence labeling it will be possible to isolate specific cell populations from *C. elegans* and subject them to metabolomics/lipidomics analysis.

6.6. Data analysis requirements

Increase in analytical throughput or deeper analysis with higher coverage or spatial analysis will require new data analysis approaches as well. Different workflows have been developed as part of this work, which are now used in routine setup that allows the reproducible analysis of the *C. elegans* metabolome and lipidome using LC-MS/MS.

Sharing of data, which includes raw will become more and more important in future. Especially data from large scale screens will represent valuable community resources. First and foremost sharing on the chemical information and reference spectra from *C. elegans* specific metabolites and detected in *C. elegans* needs to be improved. For curation of data in this work metabolites and lipids were retrieved from SI or figures and often not accompanied with machine-readable formats, such as IDs from different database or structural representations such as SMILES or InChI. Submission of data to repositories such as MetaboLights or Metabolomics Workbench will enable faster and most important automatic curation. Furthermore, spectra for species-specific metabolites such as ascarosides or others can be easier retrieved.

Data analysis and metabolite identification are often processes intimately linked to each other. Changes of metabolites can be only explained in the light of neighboring metabolites and metabolic pathways. However, unknown metabolites can be often not directly linked to pathways. Network approaches such as feature based molecular networks from GNPS hold great promises, e.g. as shown here for the analysis of daf-2 and WT worms using tandemLC. However, at the current stage this approach is only limited to metabolite features with an associated MS2. Different other network approaches have been described to link features, e.g. mass difference or correlation networks [185, 305, 511]. So far all these are typically

used individually, but each network is focusing on a specific aspect of biology. Their combination will be a powerful multilayer approach, as recently suggested in a review by Amara et al. [307]. Despite classical spectra similarity measures such as the cosine score, other measures will be used in future. The modified cosine score used by GNPS is one example, but further examples such as neutral loss spectral matching and other have been described. Furthermore, machine-learning based scores such as spec2vec will be used. In order to make full use of all possibilities, fast processing is required since typically several hundreds to thousands of spectra associated with features will be compared against each other. Different measures are available in the `Spectra` or `MetaboCoreUtils` packages, but also newer packages such `SpectriPy` enable the bridging to python and python-based approaches [369].

7. References

1. Maupas, E., Modes et formes de reproduction des nematodes. 1900.

2. Nigon, V., Les modalites de la reproduction et le determinisme due sexe chez quelques nematodes libres. Ann. Sci. Nat., 1949. **11**: p. 1-132.

3. Nigon, V., P. Guerrier, and H. Monin, L'architecture polaire de l'oeuf et les mouvements des constituants cellulaires au cours des premières étapes du développement chez quelques nématodes. Bull. Biol. Fr. Belg, 1960. **93**: p. 131-202.

4. Brenner, S., The Genetics of Caenorhabditis elegans. Genetics, 1974. **77**(1): p. 71-94.

5. Kiontke, K.C., et al., A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. BMC Evolutionary Biology, 2011. **11**(1): p. 339.

6. Andersen, E.C., et al., Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. Nature Genetics, 2012. **44**: p. 285.

7. Barrière, A. and M.-A. Félix, High Local Genetic Diversity and Low Outcrossing Rate in Caenorhabditis elegans Natural Populations. Current Biology, 2005. **15**(13): p. 1176-1184.

8. Barrière, A. and M.-A. Félix, Temporal Dynamics and Linkage Disequilibrium in Natural Caenorhabditis elegans Populations. Genetics, 2007. **176**(2): p. 999-1011.

9. Félix, M.-A. and F. Duveau, Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and C. briggsae. BMC Biology, 2012. **10**(1): p. 59.

10. Dirksen, P., et al., The native microbiome of the nematode Caenorhabditis elegans: gateway to a new host-microbiome model. BMC Biology, 2016. **14**(1): p. 38.

11. Zečić, A., I. Dhondt, and B.P. Braeckman, The nutritional requirements of Caenorhabditis elegans. Genes & Nutrition, 2019. **14**(1): p. 15.

12. Flavel, M.R., et al., Growth of Caenorhabditis elegans in Defined Media Is Dependent on Presence of Particulate Matter. G3: Genes|Genomes|Genetics, 2018. **8**(2): p. 567-575.

13. Sulston, J.E. and H.R. Horvitz, Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Developmental Biology, 1977. **56**(1): p. 110-156.

14. Kimble, J. and D. Hirsh, The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. Developmental Biology, 1979. **70**(2): p. 396-417.

15. Sulston, J.E., et al., The embryonic cell lineage of the nematode Caenorhabditis elegans. Developmental Biology, 1983. **100**(1): p. 64-119.

16. Consortium, T.C.e.S., Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science, 1998. **282**(5396): p. 2012-2018.

17. Cook, S.J., et al., Whole-animal connectomes of both Caenorhabditis elegans sexes. Nature, 2019. **571**(7763): p. 63-71.

18. Zhang, S., et al., Caenorhabditis elegans as a Useful Model for Studying Aging Mutations. Frontiers in Endocrinology, 2020. **11**.

19. Tissenbaum, H.A., Using C. elegans for aging research. Invertebr Reprod Dev, 2015. **59**(sup1): p. 59-63.

20. Thummel, C.S., Molecular Mechanisms of Developmental Timing in C. elegans and Drosophila. Developmental Cell, 2001. **1**(4): p. 453-465.

21. Kumar, A., et al., Caenorhabditis elegans: a model to understand host–microbe interactions. Cellular and Molecular Life Sciences, 2020. **77**(7): p. 1229-1249.

22. Dolgin, E.S., et al., INBREEDING AND OUTBREEDING DEPRESSION IN CAENORHABDITIS NEMATODES. Evolution, 2007. **61**(6): p. 1339-1352.

23. Fire, A., et al., Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature, 1998. **391**(6669): p. 806-811.

24. Gönczy, P., et al., Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III. Nature, 2000. **408**(6810): p. 331-336.

25. Mello, C.C. and D. Conte, Revealing the world of RNA interference. Nature, 2004. **431**(7006): p. 338-342.

26. Arribere, J.A., et al., Efficient Marker-Free Recovery of Custom Genetic Modifications with CRISPR/Cas9 in Caenorhabditis elegans. Genetics, 2014. **198**(3): p. 837-846.

27. Au, V., et al., CRISPR/Cas9 Methodology for the Generation of Knockout Deletions in Caenorhabditis elegans. G3 Genes|Genomes|Genetics, 2019. **9**(1): p. 135-144.

28. Shaye, D.D. and I. Greenwald, OrthoList: A Compendium of C. elegans Genes with Human Orthologs. PLOS ONE, 2011. **6**(5): p. e20085.

29. Kaletta, T. and M.O. Hengartner, Finding function in novel targets: C. elegans as a model organism. Nature Reviews Drug Discovery, 2006. **5**(5): p. 387-399.

30. Culetto, E. and D.B. Sattelle, A role for Caenorhabditis elegans in understanding the function and interactions of human disease genes. Human Molecular Genetics, 2000. **9**(6): p. 869-877.

31. Bürglin, T.R. and P.E. Kuwabara, Homologs of the Hh signalling network in C. elegans. WormBook.

32. Lee, R.Y N., et al., WormBase 2017: molting into a new stage. Nucleic Acids Research, 2018. **46**(D1): p. D869-D874.

33. Hulme, S.E. and G.M. Whitesides, Chemistry and the Worm: Caenorhabditis elegans as a Platform for Integrating Chemical and Biological Research. Angewandte Chemie International Edition, 2011. **50**(21): p. 4774-4807.

34. Brenner, S., Sequences and consequences. Philosophical Transactions of the Royal Society B: Biological Sciences, 2010. **365**(1537): p. 207-212.

35. Lazebnik, Y., Can a biologist fix a radio? Or, what I learned while studying apoptosis. Cancer Cell, 2002. **2**(3): p. 179-182.

36. Oliver, S.G., et al., Systematic functional analysis of the yeast genome. Trends in Biotechnology, 1998. **16**(9): p. 373-378.

37. Fiehn, O., Metabolomics – the link between genotypes and phenotypes. Plant Molecular Biology, 2002. **48**(1): p. 155-171.

38. Nicholson, J.K., 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, 1999. **29**(11): p. 1181-1189.

39. Gibson, K.M., et al., 2-Methylbutyryl-Coenzyme A Dehydrogenase Deficiency: A New Inborn Error of L-Isoleucine Metabolism. Pediatric Research, 2000. **47**(6): p. 830-833.

40. McPartland, J.M., et al., Cannabimimetic effects of osteopathic manipulative treatment. J Am Osteopath Assoc, 2005. **105**(6): p. 283-91.

41. Alonso, A., S. Marsal, and A. Julià, Analytical methods in untargeted metabolomics: state of the art in 2015. Frontiers in bioengineering and biotechnology, 2015. **3**: p. 23-23.

42. Szeto, S.S.W., S.N. Reinke, and B.D. Lemire, 1H NMR-based metabolic profiling reveals inherent biological variation in yeast and nematode model systems. Journal of Biomolecular NMR, 2011. **49**(3): p. 245-254.

43. Markley, J.L., et al., The future of NMR-based metabolomics. Current Opinion in Biotechnology, 2017. **43**: p. 34-40.

44. Whiley, L., et al., Systematic Isolation and Structure Elucidation of Urinary Metabolites Optimized for the Analytical-Scale Molecular Profiling Laboratory. Analytical Chemistry, 2019. **91**(14): p. 8873-8882.

45. Boiteau, R.M., et al., Structure Elucidation of Unknown Metabolites in Metabolomics by Combined NMR and MS/MS Prediction. Metabolites, 2018. **8**(1): p. 8.

46. Ludwig, C., et al., Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Molecular Systems Biology, 2018. **14**(8): p. e8126.

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

48. Fuhrer, T., et al., High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection–Time-of-Flight Mass Spectrometry. Analytical Chemistry, 2011. **83**(18): p. 7074-7080.

49. Pauling, L., et al., Quantitative Analysis of Urine Vapor and Breath by Gas-Liquid Partition Chromatography. Proceedings of the National Academy of Sciences, 1971. **68**(10): p. 2374-2376.

50. Buescher, J.M., et al., Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. Anal Chem, 2010. **82**(11): p. 4403-12.

51. Gao, A.W., et al., Identification of key pathways and metabolic fingerprints of longevity in C. elegans. Experimental Gerontology, 2018. **113**: p. 128-140.

52. Hastings, J., et al., Multi-Omics and Genome-Scale Modeling Reveal a Metabolic Shift During C. elegans Aging. Frontiers in Molecular Biosciences, 2019. **6**(2).

53. Xu, M., J. Legradi, and P. Leonards, Cross platform solutions to improve the zebrafish polar metabolome coverage using LC-QTOF MS: Optimization of separation mechanisms, solvent additives, and resuspension solvents. Talanta, 2021. **234**: p. 122688.

54. Haug, K., et al., MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Research, 2013. **41**(D1): p. D781-D786.

55. Sud, M., et al., Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Research, 2016. **44**(D1): p. D463-D470.

56. Harrieder, E.-M., et al., Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics. Journal of Chromatography B, 2022. **1188**: p. 123069.

57. Zhang, W. and R. Ramautar, CE-MS for metabolomics: Developments and applications in the period 2018–2020. ELECTROPHORESIS. **n/a**(n/a).

58. Drouin, N., et al., Effective mobility as a robust criterion for compound annotation and identification in metabolomics: Toward a mobility-based library. Analytica Chimica Acta, 2018. **1032**: p. 178-187.

59. Salzer, L., M. Witting, and P. Schmitt-Kopplin, MobilityTransformR: an R package for effective mobility transformation of CE-MS data. Bioinformatics, 2022. **38**(16): p. 4044-4045.

60. Fahy, E., et al., A comprehensive classification system for lipids. European Journal of Lipid Science and Technology, 2005. **107**(5): p. 337-364.

61. van Meer, G., Cellular lipidomics. EMBO J, 2005. **24**(18): p. 3159-3165.

62. Hu, C., et al., RPLC-Ion-Trap-FTMS Method for Lipid Profiling of Plasma: Method Validation and Application to p53 Mutant Mouse Model. Journal of Proteome Research, 2008. **7**(11): p. 4982-4991.

63. Wenk, M.R., The emerging field of lipidomics. Nature Reviews Drug Discovery, 2005. **4**(7): p. 594-610.

64. Surma, M.A., et al., An automated shotgun lipidomics platform for high throughput, comprehensive, and quantitative analysis of blood plasma intact lipids. European journal of lipid science and technology : EJLST, 2015. **117**(10): p. 1540-1549.

65. Wang, M. and X. Han, Multidimensional mass spectrometry-based shotgun lipidomics. Methods in molecular biology (Clifton, N.J.), 2014. **1198**: p. 203-220.

66. Schwudke, D., et al., Shotgun lipidomics on high resolution mass spectrometers. Cold Spring Harb Perspect Biol, 2011. **3**(9): p. a004614.

67. Pearson, M.J., et al., MS/MSALL with SelexION®: A High-throughput Lipidomic Solution for Untargeted Profiling. The FASEB Journal, 2020. **34**(S1): p. 1-1.

68. Witting, M., et al., Optimizing a ultrahigh pressure liquid chromatography-time of flight-mass spectrometry approach using a novel sub-2μm core–shell particle for in depth lipidomic profiling of Caenorhabditis elegans. Journal of Chromatography A, 2014. **1359**(0): p. 91-99.

69. Jankevics, A., et al., An improved strategy for analysis of lipid molecules utilising a reversed phase C30 UHPLC column and scheduled MS/MS acquisition. Talanta, 2021. **229**: p. 122262.

70. Cífková, E., et al., Lipidomic differentiation between human kidney tumors and surrounding normal tissues using HILIC-HPLC/ESI-MS and multivariate data analysis. J Chromatogr B Analyt Technol Biomed Life Sci, 2015. **1000**: p. 14-21.

71. Cífková, E., et al., Nontargeted Quantitation of Lipid Classes Using Hydrophilic Interaction Liquid Chromatography–Electrospray Ionization Mass Spectrometry with Single Internal Standard and Response Factor Approach. Analytical Chemistry, 2012. **84**(22): p. 10064-10070.

72. Lísa, M. and M. Holčapek, High-Throughput and Comprehensive Lipidomic Analysis Using Ultrahigh-Performance Supercritical Fluid Chromatography–Mass Spectrometry. Analytical Chemistry, 2015. **87**(14): p. 7187-7195.

73. Lísa, M., et al., Lipidomic analysis of biological samples: Comparison of liquid chromatography, supercritical fluid chromatography and direct infusion mass spectrometry methods. J Chromatogr A, 2017. **1525**: p. 96-108.

74. Martin, F.-P.J., et al., Metabotyping of Caenorhabditis elegans and their Culture Media Revealed Unique Metabolic Phenotypes Associated to Amino Acid Deficiency and Insulin-Like Signaling. Journal of Proteome Research, 2011. **10**(3): p. 990-1003.

75. Lourenço, A.B., et al., Analysis of the effect of the mitochondrial prohibitin complex, a context-dependent modulator of longevity, on the C. elegans metabolome. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 2015. **1847**(11): p. 1457-1468.

76. Castro, C., et al., A metabolomic strategy defines the regulation of lipid content and global metabolism by Δ9 desaturases in Caenorhabditis elegans. BMC Genomics, 2012. **13**(1): p. 36.

77. Fuchs, S., et al., A metabolic signature of long life in Caenorhabditis elegans. BMC Biology, 2010. **8**(1): p. 14.

78. Schlipalius, D.I., et al., A Core Metabolic Enzyme Mediates Resistance to Phosphine Gas. Science, 2012. **338**(6108): p. 807-810.

79. Srinivasan, J., et al., A blend of small molecules regulates both mating and development in Caenorhabditis elegans. Nature, 2008. **454**(7208): p. 1115-1118.

80. Penkov, S., et al., Maradolipids: Diacyltrehalose Glycolipids Specific to Dauer Larva in Caenorhabditis elegans. Angewandte Chemie International Edition, 2010. **49**(49): p. 9430-9435.

81. Atherton, H.J., et al., A comparative metabolomic study of NHR-49 in Caenorhabditis elegans and PPAR-α in the mouse. FEBS Letters, 2008. **582**(12): p. 1661-1666.

82. Geier, F.M., et al., Cross-Platform Comparison of Caenorhabditis elegans Tissue Extraction Strategies for Comprehensive Metabolome Coverage. Analytical Chemistry, 2011. **83**(10): p. 3730-3736.

83. An, Y.J., et al., Metabotyping of the C. elegans sir-2.1 Mutant Using in Vivo Labeling and 13C-Heteronuclear Multidimensional NMR Metabolomics. ACS Chemical Biology, 2012. **7**(12): p. 2012-2018.

84. Sheikh, M.O., et al., Correlations Between LC-MS/MS-Detected Glycomics and NMR-Detected Metabolomics in Caenorhabditis elegans Development. Frontiers in Molecular Biosciences, 2019. **6**(49).

85. Geier, F.M., A.M. Leroi, and J.G. Bundy, 13C Labeling of Nematode Worms to Improve Metabolome Coverage by Heteronuclear Nuclear Magnetic Resonance Experiments. Frontiers in Molecular Biosciences, 2019. **6**(27).

86. Wishart, D.S., et al., HMDB: the Human Metabolome Database. Nucleic Acids Research, 2007. **35**(suppl_1): p. D521-D526.

87. Wishart, D.S., et al., HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Research, 2018. **46**(D1): p. D608-D617.

88. Wishart, D.S., et al., HMDB 3.0—The Human Metabolome Database in 2013. Nucleic Acids Research, 2013. **41**(D1): p. D801-D807.

89. Wishart, D.S., et al., HMDB: a knowledgebase for the human metabolome. Nucleic Acids Research, 2009. **37**(suppl_1): p. D603-D610.

90. Mahanti, P., et al., Comparative Metabolomics Reveals Endogenous Ligands of DAF-12, a Nuclear Hormone Receptor, Regulating C. elegans Development and Lifespan. Cell metabolism, 2014. **19**(1): p. 73-83.

91. Pungaliya, C., et al., A shortcut to identifying small molecule signals that regulate behavior and development in Caenorhabditis elegans. Proceedings of the National Academy of Sciences, 2009. **106**(19): p. 7708-7713.

92. Blaise, B.J., et al., Metabotyping of Caenorhabditis elegans reveals latent phenotypes. Proceedings of the National Academy of Sciences, 2007. **104**(50): p. 19808-19812.

93. Pontoizeau, C., et al., Metabolomics Analysis Uncovers That Dietary Restriction Buffers Metabolic Changes Associated with Aging in Caenorhabditis elegans. Journal of Proteome Research, 2014. **13**(6): p. 2910-2919.

94. Wong, A., et al., µHigh Resolution-Magic-Angle Spinning NMR Spectroscopy for Metabolic Phenotyping of Caenorhabditis elegans. Analytical Chemistry, 2014. **86**(12): p. 6064-6070.

95. Sakellariou, D., G.L. Goff, and J.F. Jacquinot, High-resolution, high-sensitivity NMR of nanolitre anisotropic samples by coil spinning. Nature, 2007. **447**(7145): p. 694-697.

96. Castro, C., et al., A study of Caenorhabditis elegans DAF-2 mutants by metabolomics and differential correlation networks. Molecular bioSystems, 2013. **9**(7): p. 1632-1642.

97. Wan, Q.-L., et al., Metabolomic signature associated with reproduction-regulated aging in Caenorhabditis elegans. Aging, 2017. **9**(2): p. 447-474.

98. Teranishi, R., et al., Gas chromatography of volatiles from breath and urine. Analytical Chemistry, 1972. **44**(1): p. 18-20.

99. Jaeger, C., et al., Metabolomic changes in Caenorhabditis elegans lifespan mutants as evident from GC–EI–MS and GC–APCI–TOF–MS profiling. Metabolomics, 2014. **10**(5): p. 859-876.

100. Schwudke, D., et al., Top-Down Lipidomic Screens by Multivariate Analysis of High-Resolution Survey Mass Spectra. Analytical Chemistry, 2007. **79**(11): p. 4083-4093.

101. Lucanic, M., et al., N-acylethanolamine signalling mediates the effect of diet on lifespan in Caenorhabditis elegans. Nature, 2011. **473**(7346): p. 226-229.

102. Teo, E., et al., Metabolic stress is a primary pathogenic event in transgenic Caenorhabditis elegans expressing pan-neuronal human amyloid beta. eLife, 2019. **8**: p. e50069.

103. Folick, A., et al., Lysosomal signaling molecules regulate longevity in Caenorhabditis elegans. Science, 2015. **347**(6217): p. 83-86.

104. Wang, W., et al., Comparative Metabolomic Profiling Reveals That Dysregulated Glycolysis Stemming from Lack of Salvage NAD+ Biosynthesis Impairs Reproductive Development in Caenorhabditis elegans. Journal of Biological Chemistry, 2015. **290**(43): p. 26163-26179.

105. Henry, P., et al., Fatty acids composition of Caenorhabditis elegans using accurate mass GCMS-QTOF. Journal of environmental science and health. Part. B, Pesticides, food contaminants, and agricultural wastes, 2016. **51**(8): p. 546-552.

106. Chen, T.-L., et al., Impaired embryonic development in glucose-6-phosphate dehydrogenase-deficient Caenorhabditis elegans due to abnormal redox homeostasis induced activation of calcium-independent phospholipase and alteration of glycerophospholipid metabolism. Cell Death & Disease, 2018. **8**(1): p. e2545-e2545.

107. Mosbech, M.-B., et al., Functional Loss of Two Ceramide Synthases Elicits Autophagy-Dependent Lifespan Extension in C. elegans. PLOS ONE, 2013. **8**(7): p. e70087.

108. Klapper, M., et al., Methyl group donors abrogate adaptive responses to dietary restriction in C. elegans. Genes & Nutrition, 2016. **11**(1): p. 4.

109. Schmökel, V., et al., Genetics of Lipid-Storage Management in Caenorhabditis elegans Embryos. Genetics, 2016. **202**(3): p. 1071-1083.

110. Van Assche, R., et al., Metabolic profiling of a transgenic Caenorhabditis elegans Alzheimer model. Metabolomics : Official journal of the Metabolomic Society, 2015. **11**(2): p. 477-486.

111. Ramautar, R., G.W. Somsen, and G.J. de Jong, CE-MS for metabolomics: Developments and applications in the period 2016–2018. ELECTROPHORESIS, 2019. **40**(1): p. 165-179.

112. Blaise, B.J., et al., Metabolic Profiling Strategy of Caenorhabditis elegans by Whole-Organism Nuclear Magnetic Resonance. Journal of Proteome Research, 2009. **8**(5): p. 2542-2550.

113. Matyash, V., et al., Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. Journal of Lipid Research, 2008. **49**(5): p. 1137-1146.

114. Dall, K.B., et al., HLH-30-dependent rewiring of metabolism during starvation in C. elegans. Aging Cell. **n/a**(n/a): p. e13342.

115. Blume, B., et al., Novel Extraction Method for Combined Lipid and Metal Speciation From Caenorhabditis elegans With Focus on Iron Redox Status and Lipid Profiling. Frontiers in Chemistry, 2021. **9**.

116. Hänel, V., C. Pendleton, and M. Witting, The sphingolipidome of the model organism Caenorhabditis elegans. Chemistry and Physics of Lipids, 2019. **222**: p. 15-22.

117. Uno, M. and E. Nishida, Lifespan-regulating genes in C. elegans. npj Aging and Mechanisms of Disease, 2016. **2**(1): p. 16010.

118. Fuchs, S., et al., A metabolic signature of long life in Caenorhabditis elegans. BMC Biology, 2010. **8**(1): p. 14.

119. Davies, S.K., J.G. Bundy, and A.M. Leroi, Metabolic Youth in Middle Age: Predicting Aging in Caenorhabditis elegans Using Metabolomics. Journal of Proteome Research, 2015. **14**(11): p. 4603-4609.

120. Honda, Y., M. Tanaka, and S. Honda, Trehalose extends longevity in the nematode Caenorhabditis elegans. Aging Cell, 2010. **9**(4): p. 558-569.

121. Castro, C., et al., A study of Caenorhabditis elegans DAF-2 mutants by metabolomics and differential correlation networks. Molecular BioSystems, 2013.

122. Prasain, J.K., et al., Comparative Lipidomics of Caenorhabditis elegans Metabolic Disease Models by SWATH Non-Targeted Tandem Mass Spectrometry. Metabolites, 2015. **5**(4): p. 677-696.

123. Perez, C.L. and M.R. Van Gilst, A 13C Isotope Labeling Strategy Reveals the Influence of Insulin Signaling on Lipogenesis in C. elegans. Cell Metabolism, 2008. **8**(3): p. 266-274.

124. Depuydt, G., et al., LC–MS Proteomics Analysis of the Insulin/IGF-1-Deficient Caenorhabditis elegans daf-2(e1370) Mutant Reveals Extensive Restructuring of Intermediary Metabolism. Journal of Proteome Research, 2014. **13**(4): p. 1938-1956.

125. Butler, J.A., et al., A metabolic signature for long life in the Caenorhabditis elegans Mit mutants. Aging Cell, 2013. **12**(1): p. 130-138.

126. Mitchell, D.H., et al., Synchronous growth and aging of Caenorhabditis elegans in the presence of fluorodeoxyuridine. J Gerontol, 1979. **34**(1): p. 28-36.

127. Davies, S.K., A.M. Leroi, and J.G. Bundy, Fluorodeoxyuridine affects the identification of metabolic responses to daf-2 status in Caenorhabditis elegans. Mechanisms of Ageing and Development, 2012. **133**(1): p. 46-49.

128. Klass, M. and D. Hirsh, Non-ageing developmental variant of Caenorhabditis elegans. Nature, 1976. **260**(5551): p. 523-525.

129. Berman, J.R. and C. Kenyon, Germ-Cell Loss Extends C. elegans Life Span through Regulation of DAF-16 by kri-1 and Lipophilic-Hormone Signaling. Cell, 2006. **124**(5): p. 1055-1068.

130. Gao, A.W., et al., A sensitive mass spectrometry platform identifies metabolic changes of life history traits in C. elegans. Scientific Reports, 2017. **7**(1): p. 2408.

131. Aguilaniu, H., P. Fabrizio, and M. Witting, The Role Dafachronic Acid Signaling in Development and Longevity in Caenorhabditis elegans: Digging Deeper Using Cutting Edge Analytical Chemistry. Frontiers in Endocrinology, 2016. **7**.

132. Witting, M., et al., Fast separation and quantification of steroid hormones Δ4- and Δ7-dafachronic acid in Caenorhabditis elegans. Journal of Chromatography B, 2015. **978–979**(0): p. 118-121.

133. Thondamal, M., et al., Steroid hormone signalling links reproduction to lifespan in dietary-restricted Caenorhabditis elegans. Nat Commun, 2014. **5**.

134. Pungaliya, C., et al., A shortcut to identifying small molecule signals that regulate behavior and development in Caenorhabditis elegans. Proceedings of the National Academy of Sciences, 2009. **106**(19): p. 7708-7713.

135. Izrayelit, Y., et al., 2D NMR-Based Metabolomics Uncovers Interactions between Conserved Biochemical Pathways in the Model Organism Caenorhabditis elegans. ACS Chemical Biology, 2012.

136. von Reuss, S.H., et al., Comparative Metabolomics Reveals Biogenesis of Ascarosides, a Modular Library of Small-Molecule Signals in C. elegans. Journal of the American Chemical Society, 2012. **134**(3): p. 1817-1824.

137. Panda, O., et al., Biosynthesis of Modular Ascarosides in C. elegans. Angewandte Chemie International Edition, 2017. **56**(17): p. 4729-4733.

138. Zhou, Y., et al., Biosynthetic tailoring of existing ascaroside pheromones alters their biological function in C. elegans. eLife, 2018. **7**: p. e33286.

139. von Reuss, S.H., F. Dolke, and C. Dong, Ascaroside Profiling of Caenorhabditis elegans Using Gas Chromatography–Electron Ionization Mass Spectrometry. Analytical Chemistry, 2017. **89**(19): p. 10570-10577.

140. Artyukhin, A.B., et al., Metabolomic “Dark Matter” Dependent on Peroxisomal β-Oxidation in Caenorhabditis elegans. Journal of the American Chemical Society, 2018. **140**(8): p. 2841-2852.

141. Watts, J.L. and M. Ristow, Lipid and Carbohydrate Metabolism in Caenorhabditis elegans. Genetics, 2017. **207**(2): p. 413-446.

142. Witting, M. and P. Schmitt-Kopplin, The Caenorhabditis elegans lipidome: A primer for lipid analysis in Caenorhabditis elegans. Archives of Biochemistry and Biophysics, 2016. **589**: p. 27-37.

143. Watts, J.L. and J. Browse, Genetic dissection of polyunsaturated fatty acid synthesis in Caenorhabditis elegans. Proceedings of the National Academy of Sciences, 2002. **99**(9): p. 5854-5859.

144. Meister, P., et al., Caenorhabditis elegans Heterochromatin protein 1 (HPL-2) links developmental plasticity, longevity and lipid metabolism. Genome Biology, 2011. **12**(12): p. R123.

145. Hutzell, P.A., H. PA, and K. LR, Fatty acid compositions of Caenorhabditis elegans and C. briggsae. 1982.

146. Tanaka, T., et al., Effects of growth temperature on the fatty acid composition of the free-living nematodeCaenorhabditis elegans. Lipids, 1996. **31**(11): p. 1173-1178.

147. Brooks, K.K., B. Liang, and J.L. Watts, The Influence of Bacterial Diet on Fat Storage in C. elegans. PLOS ONE, 2009. **4**(10): p. e7545.

148. Shi, X., et al., Regulation of lipid droplet size and phospholipid composition by stearoyl-CoA desaturase. Journal of Lipid Research, 2013. **54**(9): p. 2504-2514.

149. Shi, X., et al., A Caenorhabditis elegans model for ether lipid biosynthesis and function. Journal of Lipid Research, 2016. **57**(2): p. 265-275.

150. Vrablik, T.L. and J.L. Watts, Polyunsaturated fatty acid derived signaling in reproduction and development: insights from Caenorhabditis elegans and Drosophila melanogaster. Molecular reproduction and development, 2013. **80**(4): p. 244-259.

151. Lee, H.-C., et al., Caenorhabditis elegans mboa-7, a Member of the MBOAT Family, Is Required for Selective Incorporation of Polyunsaturated Fatty Acids into Phosphatidylinositol. Molecular Biology of the Cell, 2008. **19**(3): p. 1174-1184.

152. Perez, C.L. and M.R. Van Gilst, A 13C isotope labeling strategy reveals the influence of insulin signaling on lipogenesis in C. elegans. Cell Metab, 2008. **8**(3): p. 266-74.

153. Ishida, M., et al., High-resolution analysis by nano-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for the identification of molecular species of phospholipids and their oxidized metabolites. Rapid Communications in Mass Spectrometry, 2004. **18**(20): p. 2486-2494.

154. Papan, C., et al., Systematic Screening for Novel Lipids by Shotgun Lipidomics. Analytical Chemistry, 2014. **86**(5): p. 2703-2710.

155. Witting, M., U. Schmidt, and H.-J. Knölker, UHPLC-IM-Q-ToFMS analysis of maradolipids, found exclusively in Caenorhabditis elegans dauer larvae. Analytical and Bioanalytical Chemistry, 2021. **413**(8): p. 2091-2102.

156. Menuz, V., et al., Protection of C. elegans from Anoxia by HYL-2 Ceramide Synthase. Science, 2009. **324**(5925): p. 381-384.

157. Mosbech, M.-B., et al., Functional Loss of Two Ceramide Synthases Elicits Autophagy-Dependent Lifespan Extension in <italic>C. elegans</italic>. PLoS ONE, 2013. **8**(7): p. e70087.

158. Gerdt, S., et al., Isolation, characterization and immunolocalization of phosphorylcholine-substituted glycolipids in developmental stages of Caenorhabditis elegans. European Journal of Biochemistry, 1999. **266**(3): p. 952-963.

159. Chitwood, D.J., et al., Sterol metabolism in the nematodeCaenorhabditis elegans. Lipids, 1984. **19**(7): p. 500-506.

160. Boland, S., et al., Phosphorylated glycosphingolipids essential for cholesterol mobilization in Caenorhabditis elegans. Nature Chemical Biology, 2017. **13**(6): p. 647-654.

161. Zhao, X., et al., Resolving Modifications on Sphingoid Base and N-Acyl Chain of Sphingomyelin Lipids in Complex Lipid Extracts. Analytical Chemistry, 2020. **92**(21): p. 14775-14782.

162. Drechsler, R., et al., HPLC-Based Mass Spectrometry Characterizes the Phospholipid Alterations in Ether-Linked Lipid Deficiency Models Following Oxidative Stress. PLOS ONE, 2016. **11**(11): p. e0167229.

163. Smulan, L.J., et al., Cholesterol-Independent SREBP-1 Maturation Is Linked to ARF1 Inactivation. Cell reports, 2016. **16**(1): p. 9-18.

164. Macedo, F., et al., Lipase-like 5 enzyme controls mitochondrial activity in response to starvation in Caenorhabditis elegans. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, 2020. **1865**(2): p. 158539.

165. Haeussler, S., et al., Autophagy compensates for defects in mitochondrial dynamics. PLOS Genetics, 2020. **16**(3): p. e1008638.

166. Liu, Y.J., et al., Mitochondrial translation and dynamics synergistically extend lifespan in C. elegans through HLH-30. bioRxiv, 2019: p. 871079.

167. Dall, K.B., et al., HLH-30 dependent rewiring of metabolism during starvation in C. elegans. bioRxiv, 2020: p. 2020.06.26.170555.

168. Rackles, E., et al., Reduced peroxisomal import triggers peroxisomal retrograde signaling. Cell Reports, 2021. **34**(3).

169. Admasu, T.D., et al., Drug Synergy Slows Aging and Improves Healthspan through IGF and SREBP Lipid Signaling. Developmental Cell, 2018. **47**(1): p. 67-79.e5.

170. Admasu, T.D., et al., Lipid profiling of C. elegans strains administered pro-longevity drugs and drug combinations. Scientific Data, 2018. **5**(1): p. 180231.

171. Lam, Sin M., et al., Sequestration of polyunsaturated fatty acids in membrane phospholipids of Caenorhabditis elegans dauer larva attenuates eicosanoid biosynthesis for prolonged survival. Redox Biology, 2017. **12**: p. 967-977.

172. Dancy, B.C.R., et al., 13C- and 15N-Labeling Strategies Combined with Mass Spectrometry Comprehensively Quantify Phospholipid Dynamics in C. elegans. PLOS ONE, 2015. **10**(11): p. e0141850.

173. Hughes, S.L., et al., The Metabolomic Responses of Caenorhabditis elegans to Cadmium Are Largely Independent of Metallothionein Status, but Dominated by Changes in Cystathionine and Phytochelatins. Journal of Proteome Research, 2009. **8**(7): p. 3512-3519.

174. Jones, O.A.H., et al., Potential New Method of Mixture Effects Testing Using Metabolomics and Caenorhabditis elegans. Journal of Proteome Research, 2012. **11**(2): p. 1446-1453.

175. Sudama, G., et al., Metabolic profiling in Caenorhabditis elegans provides an unbiased approach to investigations of dosage dependent lead toxicity. Metabolomics : Official journal of the Metabolomic Society, 2013. **9**(1): p. 189-201.

176. Why You Should Learn About TYPE-C Silica Columns White Paper. [cited 2020; Available from: https://cornerstonescientific.com/70000-15P-2.

177. Davies, S.K., et al., The mutational structure of metabolism in Caenorhabditis elegans. Evolution; international journal of organic evolution, 2016. **70**(10): p. 2239-2246.

178. Coburn, C., et al., Anthranilate Fluorescence Marks a Calcium-Propagated Necrotic Wave That Promotes Organismal Death in C. elegans. PLOS Biology, 2013. **11**(7): p. e1001613.

179. Blaise, B.J., et al., Metabotyping of Caenorhabditis elegans reveals latent phenotypes. Proceedings of the National Academy of Sciences, 2007. **104**(50): p. 19808-19812.

180. Atherton, H.J., et al., A comparative metabolomic study of NHR-49 in Caenorhabditis elegans and PPAR-[alpha] in the mouse. FEBS Letters, 2008. **582**(12): p. 1661-1666.

181. Hughes, S.L., et al., The Metabolomic Responses of Caenorhabditis elegans to Cadmium Are Largely Independent of Metallothionein Status, but Dominated by Changes in Cystathionine and Phytochelatins. Journal of Proteome Research, 2009. **8**(7): p. 3512-3519.

182. Jones, O.A.H., et al., Potential New Method of Mixture Effects Testing Using Metabolomics and Caenorhabditis elegans. Journal of Proteome Research, 2011. **11**(2): p. 1446-1453.

183. Butler, J.A., et al., Profiling the Anaerobic Response of <italic>C. elegans</italic> Using GC-MS. PLoS ONE, 2012. **7**(9): p. e46140.

184. Folick, A., et al., Lysosomal signaling molecules regulate longevity in Caenorhabditis elegans. Science, 2015. **347**(6217): p. 83-86.

185. Witting, M., et al., DI-ICR-FT-MS-based high-throughput deep metabotyping: a case study of the Caenorhabditis elegans–Pseudomonas aeruginosa infection model. Analytical and Bioanalytical Chemistry, 2015. **407**(4): p. 1059-1073.

186. Witting, M., et al., LipidFrag: Improving reliability of in silico fragmentation of lipids and application to the Caenorhabditis elegans lipidome. PLOS ONE, 2017. **12**(3): p. e0172311.

187. Dall, K.B., et al., HLH-30-dependent rewiring of metabolism during starvation in C. elegans. Aging Cell, 2021. **20**(4): p. e13342.

188. Bensaddek, D., et al., Micro-proteomics with iterative data analysis: Proteome analysis in C. elegans at the single worm level. PROTEOMICS, 2016. **16**(3): p. 381-392.

189. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res, 2021. **49**(D1): p. D480-d489.

190. Kanehisa, M., et al., KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research, 2017. **45**(D1): p. D353-D361.

191. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000. **28**.

192. Kanehisa, M., et al., From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Research, 2006. **34**(suppl 1): p. D354-D357.

193. Kanehisa, M., et al., Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res, 2014. **42**.

194. Suhre, K. and P. Schmitt-Kopplin, MassTRIX: mass translator into pathways. Nucleic Acids Research, 2008. **36**(suppl 2): p. W481-W484.

195. Wägele, B., et al., MassTRIX Reloaded: Combined Analysis and Visualization of Transcriptome and Metabolome Data. PLoS One, 2012. **7**(7): p. e39860.

196. Witting, M. and P. Schmitt-Kopplin, Chapter 17 - Transcriptome and Metabolome Data Integration—Technical Perquisites for Successful Data Fusion and Visualization, in Comprehensive Analytical Chemistry, A.C. Carolina Simó and G.-C. Virginia, Editors. 2014, Elsevier. p. 421-442.

197. García-Alcalde, F., et al., Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics, 2011. **27**(1): p. 137-139.

198. Wishart, D.S., et al., HMDB 3.0—the human metabolome database in 2013. Nucleic Acids Res, 2013. **41**.

199. Wishart, D.S., et al., HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res, 2009(37 Database): p. D603-610.

200. Ramirez-Gaona, M., et al., YMDB 2.0: a significantly expanded version of the yeast metabolome database. Nucleic Acids Res, 2017. **45**(D1): p. D440-d445.

201. Sajed, T., et al., ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. Nucleic Acids Res, 2016. **44**(D1): p. D495-501.

202. Hastings, J., et al., The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Research, 2013. **41**(D1): p. D456-D463.

203. Hastings, J., et al., ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res, 2016. **44**(D1): p. D1214-9.

204. Poupin, N., et al., Improving lipid mapping in Genome Scale Metabolic Networks using ontologies. Metabolomics, 2020. **16**(4): p. 44.

205. Fahy, E., et al., Update of the LIPID MAPS comprehensive classification system for lipids. Journal of Lipid Research, 2009. **50**(Suppl): p. S9-S14.

206. Bridge, A., et al., The SwissLipids knowledgebase for lipid biology. Bioinformatics, 2015. **31**(17): p. 2860-2866.

207. Liebisch, G., et al., Shorthand notation for lipid structures derived from mass spectrometry. Journal of Lipid Research, 2013. **54**(6): p. 1523-1530.

208. Chang, A., et al., BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res, 2021. **49**(D1): p. D498-d508.

209. Bansal, P., et al., Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res, 2022. **50**(D1): p. D693-d700.

210. Martens, M., et al., WikiPathways: connecting communities. Nucleic Acids Research, 2020. **49**(D1): p. D613-D621.

211. Gillespie, M., et al., The reactome pathway knowledgebase 2022. Nucleic Acids Res, 2022. **50**(D1): p. D687-d692.

212. Caspi, R., et al., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research, 2016. **44**(D1): p. D471-D480.

213. Gebauer, J., et al., A Genome-Scale Database and Reconstruction of Caenorhabditis elegans Metabolism. Cell Systems, 2016. **2**(5): p. 312-322.

214. Wishart, D.S., et al., PathBank: a comprehensive pathway database for model organisms. Nucleic Acids Research, 2019. **48**(D1): p. D470-D478.

215. Brunk, E., et al., Recon3D enables a three-dimensional view of gene variation in human metabolism. Nature Biotechnology, 2018. **36**: p. 272.

216. Khodaee, S., et al., iMM1865: A New Reconstruction of Mouse Genome-Scale Metabolic Model. Scientific Reports, 2020. **10**(1): p. 6177.

217. Lu, H., et al., A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nature Communications, 2019. **10**(1): p. 3586.

218. Edwards, J.S. and B.O. Palsson, Systems Properties of the Haemophilus influenzaeRd Metabolic Genotype. Journal of Biological Chemistry, 1999. **274**(25): p. 17410-17416.

219. Krauss, M., et al., Integrating Cellular Metabolism into a Multiscale Whole-Body Model. PLOS Computational Biology, 2012. **8**(10): p. e1002750.

220. Yilmaz, L.S. and A.J.M. Walhout, Metabolic network modeling with model organisms. Current Opinion in Chemical Biology, 2017. **36**: p. 32-39.

221. Karp, P.D., S. Paley, and P. Romero, The Pathway Tools software. Bioinformatics, 2002. **18 Suppl 1**: p. S225-32.

222. Yilmaz, L.S. and Albertha J.M. Walhout, A Caenorhabditis elegans Genome-Scale Metabolic Network Model. Cell Systems, 2016. **2**(5): p. 297-311.

223. Metabolic Modeling for Design of Cell Factories, in Systems Biology.

224. Orth, J.D., I. Thiele, and B.Ø. Palsson, What is flux balance analysis? Nature Biotechnology, 2010. **28**(3): p. 245-248.

225. King, Z.A., et al., BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Research, 2015. **44**(D1): p. D515-D522.

226. Cottret, L., et al., MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res, 2018. **46**(W1): p. W495-w502.

227. Büchel, F., et al., Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Systems Biology, 2013. **7**(1): p. 116.

228. Ma, L., et al., Systems Biology Analysis Using A Genome-Scale Metabolic Model Shows That Phosphine Triggers Global Metabolic Suppression In A Resistant Strain Of C. elegans. bioRxiv, 2017.

229. Wang, H., et al., Genome-scale metabolic network reconstruction of model animals as a platform for translational research. Proceedings of the National Academy of Sciences, 2021. **118**(30): p. e2102344118.

230. Hastings, J., et al., WormJam: A consensus C. elegans Metabolic Reconstruction and Metabolomics Community and Workshop Series. Worm, 2017. **6**(2): p. e1373939.

231. Witting, M., et al., Modeling Meets Metabolomics—The WormJam Consensus Model as Basis for Metabolic Studies in the Model Organism Caenorhabditis elegans. Frontiers in Molecular Biosciences, 2018. **5**(96).

232. King, Z.A., et al., BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Research, 2016. **44**(D1): p. D515-D522.

233. Moretti, S., et al., MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Research, 2016. **44**(D1): p. D523-D526.

234. van Heck, R.G.A., et al., Efficient Reconstruction of Predictive Consensus Metabolic Network Models. PLOS Computational Biology, 2016. **12**(8): p. e1005085.

235. Hucka, M., et al., The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics, 2003. **19**(4): p. 524-531.

236. Hucka, M., et al., Systems Biology Markup Language (SBML) Level 2 Version 5: Structures and Facilities for Model Definitions. Journal of integrative bioinformatics, 2015. **12**(2): p. 271-271.

237. Hucka, M., et al., The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core, in Journal of Integrative Bioinformatics. 2018.

238. Lubitz, T., et al., SBtab: a flexible table format for data exchange in systems biology. Bioinformatics, 2016. **32**(16): p. 2559-2561.

239. Lieven, C., et al., MEMOTE for standardized genome-scale metabolic model testing. Nature Biotechnology, 2020. **38**(3): p. 272-276.

240. Moretti, S., et al., MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Research, 2015. **44**(D1): p. D523-D526.

241. Ravikrishnan, A. and K. Raman, Critical assessment of genome-scale metabolic networks: the need for a unified standard. Briefings in Bioinformatics, 2015. **16**(6): p. 1057-1068.

242. Carey, M.A., et al., Community standards to facilitate development and address challenges in metabolic modeling. bioRxiv, 2019: p. 700112.

243. Lieven, C., et al., Memote: A community driven effort towards a standardized genome-scale metabolic model test suite. bioRxiv, 2018: p. 350991.

244. van Iersel, M.P., et al., The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics, 2010. **11**(1): p. 5.

245. van Rijswijk, M., et al., The future of metabolomics in ELIXIR [version 2; peer review: 3 approved]. F1000Research, 2017. **6**(1649).

246. Brunk, E., et al., Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat Biotechnol, 2018. **36**(3): p. 272-281.

247. Liebisch, G., et al., Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures. Journal of Lipid Research, 2020. **61**(12): p. 1539-1555.

248. Hefzi, H., et al., A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. Cell Systems, 2016. **3**(5): p. 434-443.e8.

249. Bulcha, J.T., et al., A Persistence Detector for Metabolic Network Rewiring in an Animal. Cell Reports, 2019. **26**(2): p. 460-468.e4.

250. Zhu, H., et al., A novel sphingolipid-TORC1 pathway critically promotes postembryonic development in Caenorhabditis elegans, ed. J. Ahringer. Vol. 2. 2013.

251. Hannich, J.T., et al., Structure and conserved function of iso-branched sphingoid bases from the nematode Caenorhabditis elegans. Chemical Science, 2017. **8**(5): p. 3676-3686.

252. Watts, J.L. and J. Browse, Genetic dissection of polyunsaturated fatty acid synthesis in Caenorhabditis elegans. Proceedings of the National Academy of Sciences, 2002. **99**(9): p. 5854-5859.

253. Gurvitz, A., A C. elegans Model for Mitochondrial Fatty Acid Synthase II: The Longevity-Associated Gene W09H1.5/mecr-1 Encodes a 2-trans-Enoyl-Thioester Reductase. PLOS ONE, 2009. **4**(11): p. e7791.

254. Kniazeva, M., et al., Monomethyl Branched-Chain Fatty Acids Play an Essential Role in Caenorhabditis elegans Development. PLoS Biol, 2004. **2**(9): p. e257.

255. Zhang, H., et al., Apicobasal domain identities of expanding tubular membranes depend on glycosphingolipid biosynthesis. Nature Cell Biology, 2011. **13**: p. 1189.

256. Entchev, E.V., et al., LET-767 Is Required for the Production of Branched Chain and Long Chain Fatty Acids in Caenorhabditis elegans. Journal of Biological Chemistry, 2008. **283**(25): p. 17550-17560.

257. Watts, J.L. and J. Browse, A Palmitoyl-CoA-Specific Δ9 Fatty Acid Desaturase from Caenorhabditis elegans. Biochemical and Biophysical Research Communications, 2000. **272**(1): p. 263-269.

258. Brock, T.J., J. Browse, and J.L. Watts, Fatty Acid Desaturation and the Regulation of Adiposity in Caenorhabditis elegans. Genetics, 2007. **176**(2): p. 865-875.

259. Hoang, H.D., et al., A heterogeneous mixture of F-series prostaglandins promotes sperm guidance in the Caenorhabditis elegans reproductive tract. PLoS genetics, 2013. **9**(1): p. e1003271-e1003271.

260. Edmonds, J.W., et al., Insulin/FOXO Signaling Regulates Ovarian Prostaglandins Critical for Reproduction. Developmental Cell, 2010. **19**(6): p. 858-871.

261. Tiwary, E., et al., Signature profile of cyclooxygenase-independent F2 series prostaglandins in C. elegans and their role in sperm motility. Scientific Reports, 2019. **9**(1): p. 11750.

262. Guo, L., et al., Stereospecificity of fatty acid 2-hydroxylase and differential functions of 2-hydroxy fatty acid enantiomers. Journal of Lipid Research, 2012. **53**(7): p. 1327-1335.

263. Zhang, X., et al., Acyl-CoA oxidase complexes control the chemical message produced by Caenorhabditis elegans. Proceedings of the National Academy of Sciences, 2015. **112**(13): p. 3955-3960.

264. Gao, A.W., et al., Natural genetic variation in C. elegans identified genomic loci controlling metabolite levels. Genome Research, 2018. **28**(9): p. 1296-1308.

265. Butcher, R.A., J.R. Ragains, and J. Clardy, An Indole-Containing Dauer Pheromone Component with Unusual Dauer Inhibitory Activity at Higher Concentrations. Organic Letters, 2009. **11**(14): p. 3100-3103.

266. Srinivasan, J., et al., A Modular Library of Small Molecule Signals Regulates Social Behaviors in Caenorhabditis elegans. PLOS Biology, 2012. **10**(1): p. e1001237.

267. Merris, M., et al., Sterol effects and sites of sterol accumulation in Caenorhabditis elegans: developmental requirement for 4α-methyl sterols. Journal of Lipid Research, 2003. **44**(1): p. 172-181.

268. Wollam, J., et al., The Rieske oxygenase DAF-36 functions as a cholesterol 7-desaturase in steroidogenic pathways governing longevity. Aging Cell, 2011. **10**(5): p. 879-884.

269. Wang, W., et al., Comparative Metabolomic Profiling Reveals That Dysregulated Glycolysis Stemming from Lack of Salvage NAD+ Biosynthesis Impairs Reproductive Development in Caenorhabditis elegans. Journal of Biological Chemistry, 2015. **290**(43): p. 26163-26179.

270. Lands, W.E.M., METABOLISM OF GLYCEROLIPIDES: A COMPARISON OF LECITHIN AND TRIGLYCERIDE SYNTHESIS. Journal of Biological Chemistry, 1958. **231**(2): p. 883-888.

271. Matsuda, S., et al., Member of the membrane-bound O-acyltransferase (MBOAT) family encodes a lysophospholipid acyltransferase with broad substrate specificity. Genes to Cells, 2008. **13**(8): p. 879-888.

272. Ye, C., Z. Shen, and M.L. Greenberg, Cardiolipin remodeling: a regulatory hub for modulating cardiolipin metabolism and function. Journal of bioenergetics and biomembranes, 2016. **48**(2): p. 113-123.

273. Solari, F., et al., The human tumour suppressor PTEN regulates longevity and dauer formation in Caenorhabditis elegans. Oncogene, 2005. **24**(1): p. 20-7.

274. Penkov, S., OAC-39 is an O-acyltransferase required for the synthesis of maradolipids in the dauer larva of C. elegans. bioRxiv, 2021: p. 2021.10.15.464527.

275. Harrison, N., et al., Characterization of N-Acyl Phosphatidylethanolamine-Specific Phospholipase-D Isoforms in the Nematode Caenorhabditis elegans. PLOS ONE, 2014. **9**(11): p. e113007.

276. Pastuhov, S.I., K. Matsumoto, and N. Hisamoto, Endocannabinoid signaling regulates regenerative axon navigation in Caenorhabditis elegans via the GPCRs NPR-19 and NPR-32. Genes to Cells, 2016. **21**(7): p. 696-705.

277. Prasain, J., et al., Comparative Lipidomics of Caenorhabditis elegans Metabolic Disease Models by SWATH Non-Targeted Tandem Mass Spectrometry. Metabolites, 2015. **5**(4): p. 677.

278. Maccarrone, M., Metabolism of the Endocannabinoid Anandamide: Open Questions after 25 Years. Frontiers in Molecular Neuroscience, 2017. **10**(166).

279. Chitwood, D., et al., The glycosylceramides of the nematodeCaenorhabditis elegans contain an unusual, branched-chain sphingoid base. Lipids, 1995. **30**(6): p. 567-573.

280. Gerdt, S., et al., Isolation and structural analysis of three neutral glycosphingolipids from a mixed population of Caenorhabditis elegans (Nematoda: Rhabditida). Glycobiology, 1997. **7**(2): p. 265-275.

281. Sumner, L., et al., Proposed minimum reporting standards for chemical analysis. Metabolomics, 2007. **3**(3): p. 211-221.

282. Sánchez, B.J., et al., SLIMEr: probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework. BMC Systems Biology, 2019. **13**(1): p. 4.

283. Biazzo, I., A. Braunstein, and R. Zecchina, Performance of a cavity-method-based algorithm for the prize-collecting Steiner tree problem on graphs. Physical Review E, 2012. **86**(2): p. 026706.

284. Pirhaji, L., et al., Revealing disease-associated pathways by network integration of untargeted metabolomics. Nature Methods, 2016. **13**: p. 770.

285. Nguyen, A., et al., Host lipidome analysis during rhinovirus replication in HBECs identifies potential therapeutic targets. Journal of Lipid Research, 2018. **59**(9): p. 1671-1684.

286. Köhler, N., et al., Investigating Global Lipidome Alterations with the Lipid Network Explorer. Metabolites, 2021. **11**(8): p. 488.

287. Nguyen, A., et al., Using lipidomics analysis to determine signalling and metabolic changes in cells. Current Opinion in Biotechnology, 2017. **43**: p. 96-103.

288. Peng, B., et al., LipidCreator workbench to probe the lipidomic landscape. Nature Communications, 2020. **11**(1): p. 2057.

289. Aimo, L., et al., The SwissLipids knowledgebase for lipid biology. Bioinformatics, 2015. **31**(17): p. 2860-2866.

290. Vrablik, T.L., et al., Lipidomic and proteomic analysis of Caenorhabditis elegans lipid droplets and identification of ACS-4 as a lipid droplet-associated protein. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, 2015. **1851**(10): p. 1337-1345.

291. Tanaka, T., et al., Effects of growth temperature on the fatty acid composition of the free-living nematodeCaenorhabditis elegans. Lipids, 1996. **31**(11): p. 1173-1178.

292. Hutzell, P.A. and L. Krusberg, Fatty acid compositions of Caenorhabditis elegans and C. briggsae. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry, 1982. **73**(3): p. 517-520.

293. Kopczynski, D., et al., Goslin: A Grammar of Succinct Lipid Nomenclature. Analytical Chemistry, 2020. **92**(16): p. 10957-10960.

294. Kopczynski, D., et al., Goslin 2.0 Implements the Recent Lipid Shorthand Nomenclature for MS-Derived Lipid Structures. Analytical Chemistry, 2022. **94**(16): p. 6097-6101.

295. Helmer, P.O., et al., Investigation of cardiolipin oxidation products as a new endpoint for oxidative stress in C. elegans by means of online two-dimensional liquid chromatography and high-resolution mass spectrometry. Free Radical Biology and Medicine, 2021. **162**: p. 216-224.

296. Oemer, G., et al., Molecular structural diversity of mitochondrial cardiolipins. Proceedings of the National Academy of Sciences, 2018. **115**(16): p. 4158-4163.

297. Schymanski, E.L., et al., Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environmental Science & Technology, 2014. **48**(4): p. 2097-2098.

298. Wishart, D.S., et al., HMDB 3.0—The Human Metabolome Database in 2013. Nucleic Acids Research, 2012.

299. Horai, H., et al., MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom, 2010. **45**.

300. Wang, M., et al., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature Biotechnology, 2016. **34**(8): p. 828-837.

301. Xue, J., et al., METLIN MS2 molecular standards database: a broad chemical and biological resource. Nature Methods, 2020. **17**(10): p. 953-954.

302. Nagao, T., et al., Power of isotopic fine structure for unambiguous determination of metabolite elemental compositions: In silico evaluation and metabolomic application. Analytica Chimica Acta, 2014. **813**: p. 70-76.

303. Kind, T. and O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics, 2007. **8**(1): p. 105.

304. Böcker, S., et al., SIRIUS: decomposing isotope patterns for metabolite identification†. Bioinformatics, 2009. **25**(2): p. 218-224.

305. Tziotis, D., Norbert Hertkorn, and P. Schmitt-Kopplin, Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. European Journal of Mass Spectrometry, 2011. **17**(4): p. 6.

306. Naake, T. and A.R. Fernie, MetNet: Metabolite Network Prediction from High-Resolution Mass Spectrometry Data in R Aiding Metabolite Annotation. Analytical Chemistry, 2019. **91**(3): p. 1768-1772.

307. Amara, A., et al., Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. Frontiers in Molecular Biosciences, 2022. **9**.

308. Rasche, F., et al., Identifying the Unknowns by Aligning Fragmentation Trees. Analytical Chemistry, 2012. **84**(7): p. 3417-3426.

309. Ludwig, M., et al., Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. Nature Machine Intelligence, 2020. **2**(10): p. 629-641.

310. Smith, C.A., et al., METLIN: a metabolite mass spectral database. Ther Drug Monit, 2005. **27**(6): p. 747-751.

311. Guijas, C., et al., METLIN: A Technology Platform for Identifying Knowns and Unknowns. Analytical Chemistry, 2018. **90**(5): p. 3156-3164.

312. Ruttkies, C., et al., MetFrag relaunched: incorporating strategies beyond in silico fragmentation. Journal of Cheminformatics, 2016. **8**(1): p. 3.

313. Dührkop, K., et al., Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci, 2015. **112**.

314. Djoumbou-Feunang, Y., et al., CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. Metabolites, 2019. **9**(4): p. 72.

315. Schymanski, E.L., et al., Critical Assessment of Small Molecule Identification 2016: automated methods. Journal of Cheminformatics, 2017. **9**(1): p. 22.

316. Hoffmann, M.A., et al., High-confidence structural annotation of metabolites absent from spectral libraries. Nature Biotechnology, 2022. **40**(3): p. 411-421.

317. Watrous, J., et al., Mass spectral molecular networking of living microbial colonies. Proceedings of the National Academy of Sciences, 2012. **109**(26): p. E1743-E1752.

318. Ernst, M., et al., MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. Metabolites, 2019. **9**(7): p. 144.

319. Müller, C., et al., Molecular cartography in acute Chlamydia pneumoniae infections—a non-targeted metabolomics approach. Analytical and Bioanalytical Chemistry, 2013. **405**(15): p. 5119-5131.

320. Stanstrup, J., et al., Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Analytical and Bioanalytical Chemistry, 2013. **405**(15): p. 5037-5048.

321. Yu, Z., et al., Optimizing 2D gas chromatography mass spectrometry for robust tissue, serum and urine metabolite profiling. Talanta, 2017. **165**: p. 685-691.

322. Lipok, C., J. Hippler, and O.J. Schmitz, A four dimensional separation method based on continuous heart-cutting gas chromatography with ion mobility and high resolution mass spectrometry. Journal of chromatography. A, 2018. **1536**: p. 50-57.

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

324. Abate-Pella, D., et al., Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods. Journal of Chromatography A, 2015. **1412**: p. 43-51.

325. Héberger, K., Quantitative structure–(chromatographic) retention relationships. Journal of Chromatography A, 2007. **1158**(1): p. 273-305.

326. Cheng, T., et al., Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. Journal of Chemical Information and Modeling, 2007. **47**(6): p. 2140-2148.

327. Alpert, A.J., Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. Journal of Chromatography A, 1990. **499**(0): p. 177-196.

328. Kaliszan, R., QUANTITATIVE STRUCTURE-RETENTION RELATIONSHIPS. Analytical Chemistry, 1992. **64**(11): p. 619A-631A.

329. Cao, M., et al., Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. Metabolomics, 2015. **11**.

330. Hagiwara, T., et al., HPLC Retention time prediction for metabolome analysis Bioinformation, 2010. **5**(6).

331. Kohlbacher, O., et al., Structure–Activity Relationships in Chromatography: Retention Prediction of Oligonucleotides with Support Vector Regression. Angewandte Chemie International Edition, 2006. **45**(42): p. 7009-7012.

332. Aalizadeh, R., M.-C. Nika, and N.S. Thomaidis, Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. Journal of Hazardous Materials, 2019. **363**: p. 277-285.

333. Mollerup, C.B., et al., Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry. Journal of Chromatography A, 2018. **1542**: p. 82-88.

334. Eugster, P.J., et al., Retention time prediction for dereplication of natural products (CxHyOz) in LC–MS metabolite profiling. Phytochemistry, 2014. **108**: p. 196-207.

335. Bączek, T. and R. Kaliszan, Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. PROTEOMICS, 2009. **9**(4): p. 835-847.

336. Krokhin, O.V. and V. Spicer, Peptide Retention Standards and Hydrophobicity Indexes in Reversed-Phase High-Performance Liquid Chromatography of Peptides. Analytical Chemistry, 2009. **81**(22): p. 9522-9530.

337. Ma, C., et al., Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. Analytical Chemistry, 2018. **90**(18): p. 10881-10888.

338. Dorfer, V., et al., CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. Journal of Proteome Research, 2018. **17**(8): p. 2581-2589.

339. Aicheler, F., et al., Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches. Analytical Chemistry, 2015. **87**(15): p. 7698-7704.

340. Tsugawa, H., et al., Comprehensive identification of sphingolipid species by in silico retention time and tandem mass spectral library. Journal of Cheminformatics, 2017. **9**(1): p. 19.

341. Goudarzi, N., et al., Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons. Journal of Chromatography A, 2014. **1333**(0): p. 25-31.

342. Creek, D.J., et al., Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. Analytical Chemistry, 2011. **83**(22): p. 8703-8710.

343. Randazzo, G.M., et al., Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. Analytica Chimica Acta, 2016. **916**: p. 8-16.

344. Jaramillo, R. and F.L. Dorman, Retention time prediction in thermally modulated comprehensive two-dimensional gas chromatography: Correcting second dimension retention time modeling error. Journal of Chromatography A, 2018.

345. Veenaas, C., A. Linusson, and P. Haglund, Retention-time prediction in comprehensive two-dimensional gas chromatography to aid identification of unknown contaminants. Analytical and Bioanalytical Chemistry, 2018. **410**(30): p. 7931-7941.

346. Burel, A., et al., Retention modeling and retention time prediction in gas chromatography and flow-modulation comprehensive two-dimensional gas chromatography: The contribution of pressure on solute partition. Journal of Chromatography A, 2017. **1485**: p. 101-119.

347. McEachran, A.D., et al., A comparison of three liquid chromatography (LC) retention time prediction models. Talanta, 2018. **182**: p. 371-379.

348. Colby, S.M., et al., ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries. Analytical Chemistry, 2019. **91**(7): p. 4346-4356.

349. Colby, S.M., et al., Deep Learning to Generate in Silico Chemical Property Libraries and Candidate Molecules for Small Molecule Identification in Complex Samples. Analytical Chemistry, 2020. **92**(2): p. 1720-1729.

350. Ross, D.H., J.H. Cho, and L. Xu, Breaking Down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections. Analytical Chemistry, 2020. **92**(6): p. 4548-4557.

351. Zhou, Z., X. Xiong, and Z.-J. Zhu, MetCCS predictor: a web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. Bioinformatics, 2017. **33**(14): p. 2235-2237.

352. Poad, B.L.J., et al., Ozone-Induced Dissociation on a Modified Tandem Linear Ion-Trap: Observations of Different Reactivity for Isomeric Lipids. Journal of the American Society for Mass Spectrometry, 2010. **21**(12): p. 1989-1999.

353. Griss, J., et al., The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. Molecular & Cellular Proteomics, 2014. **13**(10): p. 2765-2775.

354. Cabruja, M., et al., In-depth triacylglycerol profiling using MS3 Q-Trap mass spectrometry. Analytica Chimica Acta, 2021. **1184**: p. 339023.

355. Kind, T., M. Scholz, and O. Fiehn, How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry. PLoS ONE, 2009. **4**(5): p. e5440.

356. Wang, M., et al., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature Biotechnology, 2016. **34**: p. 828.

357. da Silva, R.R., et al., Propagating annotations of molecular networks using in silico fragmentation. PLOS Computational Biology, 2018. **14**(4): p. e1006089.

358. Kováts, E., Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. Helv Chim Acta, 1958. **41**.

359. Strehmel, N., et al., Retention index thresholds for compound matching in GC–MS metabolite profiling. Journal of Chromatography B, 2008. **871**(2): p. 182-190.

360. Aderjan, R. and M. Bogusz, Nitroalkanes as a multidetector retention index scale for drug identification in gas chromatography. Journal of Chromatography A, 1988. **454**: p. 345-351.

361. Smith, R.M., Chapter 3 Retention index scales used in high-performance liquid chromatography, in Journal of Chromatography Library, R.M. Smith, Editor. 1995, Elsevier. p. 93-144.

362. Hall, L.M., et al., Development of Ecom(50) and Retention Index Models for Non-Targeted Metabolomics: Identification of 1,3-dicyclohexylurea in Human Serum by HPLC/Mass Spectrometry. Journal of Chemical Information and Modeling, 2012. **52**(5): p. 1222-1237.

363. Allen, F., R. Greiner, and D. Wishart, Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics, 2015. **11**(1): p. 98-110.

364. Wolf, S., et al., In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinform, 2010. **11**.

365. Samaraweera, M.A., et al., Evaluation of an Artificial Neural Network Retention Index Model for Chemical Structure Identification in Nontargeted Metabolomics. Analytical Chemistry, 2018.

366. Zheng, S.-J., et al., Establishment of Liquid Chromatography Retention Index Based on Chemical Labeling for Metabolomic Analysis. Analytical Chemistry, 2018. **90**(14): p. 8412-8420.

367. Renaud, J.B., et al., Normalization of LC-MS mycotoxin determination using the N-alkylpyridinium-3-sulfonates (NAPS) retention index system. Journal of Chromatography A, 2021. **1639**: p. 461901.

368. Team, R.C., R: A Language and Environment for Statistical Computing. 2012.

369. Rainer, J., et al., A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. Metabolites, 2022. **12**(2): p. 173.

370. Akima, H., et al., akima: Interpolation of Irregularly and Regularly Spaced Data. 2021.

371. QUILLIAM, M.A., Retention index standards for liquid chromatography. 2013, Google Patents.

372. Snyder, L.R., J.W. Dolan, and P.W. Carr, The hydrophobic-subtraction model of reversed-phase column selectivity. Journal of Chromatography A, 2004. **1060**(1): p. 77-116.

373. Pezzatti, J., et al., A scoring approach for multi-platform acquisition in metabolomics. Journal of Chromatography A, 2019. **1592**: p. 47-54.

374. Paglia, G., et al., Ion Mobility Derived Collision Cross Sections to Support Metabolomics Applications. Analytical Chemistry, 2014. **86**(8): p. 3985-3993.

375. Zheng, X., et al., A structural examination and collision cross section database for over 500 metabolites and xenobiotics using drift tube ion mobility spectrometry. Chemical Science, 2017. **8**(11): p. 7724-7736.

376. Stow, S.M., et al., An Interlaboratory Evaluation of Drift Tube Ion Mobility–Mass Spectrometry Collision Cross Section Measurements. Analytical Chemistry, 2017. **89**(17): p. 9048-9055.

377. Nye, L.C., et al., A comparison of collision cross section values obtained via travelling wave ion mobility-mass spectrometry and ultra high performance liquid chromatography-ion mobility-mass spectrometry: Application to the characterisation of metabolites in rat urine. Journal of Chromatography A, 2019. **1602**: p. 386-396.

378. Hinnenkamp, V., et al., Comparison of CCS Values Determined by Traveling Wave Ion Mobility Mass Spectrometry and Drift Tube Ion Mobility Mass Spectrometry. Analytical Chemistry, 2018. **90**(20): p. 12042-12050.

379. Picache, J.A., et al., Collision cross section compendium to annotate and predict multi-omic compound identities. Chemical Science, 2019. **10**(4): p. 983-993.

380. Zhou, Z., et al., Ion mobility collision cross-section atlas for known and unknown metabolite annotation in untargeted metabolomics. Nature Communications, 2020. **11**(1): p. 4334.

381. Leaptrot, K.L., et al., Ion mobility conformational lipid atlas for high confidence lipidomics. Nature Communications, 2019. **10**(1): p. 985.

382. Tsugawa, H., et al., A lipidome atlas in MS-DIAL 4. Nature Biotechnology, 2020.

383. Pässler, U., et al., Synthesis of Ten Members of the Maradolipid Family; Novel Diacyltrehalose Glycolipids from Caenorhabditis elegans. Synlett, 2011. **2011**(17): p. 2482-2486.

384. Bligh, E.G. and W.J. Dyer, A rapid method of total lipid extraction and purification. Canadian Journal of Biochemistry and Physiology, 1959. **37**(8): p. 911-917.

385. Knittelfelder, O.L., et al., A versatile ultra-high performance LC-MS method for lipid profiling. Journal of Chromatography B, 2014. **951-952**: p. 119-128.

386. Schymanski, E.L., et al., Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. Environ Sci Technol, 2014. **48**.

387. Causon, T.J. and S. Hann, Theoretical evaluation of peak capacity improvements by use of liquid chromatography combined with drift tube ion mobility-mass spectrometry. Journal of Chromatography A, 2015. **1416**: p. 47-56.

388. Willighagen, E.L., et al., The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics, 2017. **9**(1): p. 33.

389. Mesleh, M.F., et al., Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential. The Journal of Physical Chemistry, 1996. **100**(40): p. 16082-16086.

390. Shvartsburg, A.A. and M.F. Jarrold, An exact hard-spheres scattering model for the mobilities of polyatomic ions. Chemical Physics Letters, 1996. **261**(1): p. 86-91.

391. Ieritano, C., et al., A parallelized molecular collision cross section package with optimized accuracy and efficiency. Analyst, 2019. **144**(5): p. 1660-1670.

392. Zhou, Z., et al., Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. Analytical Chemistry, 2016. **88**(22): p. 11084-11091.

393. Zhou, Z., et al., LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility–Mass Spectrometry-Based Lipidomics. Analytical Chemistry, 2017. **89**(17): p. 9559-9566.

394. Hines, K.M., et al., Evaluation of Collision Cross Section Calibrants for Structural Analysis of Lipids by Traveling Wave Ion Mobility-Mass Spectrometry. Analytical Chemistry, 2016. **88**(14): p. 7329-7336.

395. Blaženović, I., et al., Increasing Compound Identification Rates in Untargeted Lipidomics Research with Liquid Chromatography Drift Time–Ion Mobility Mass Spectrometry. Analytical Chemistry, 2018. **90**(18): p. 10758-10764.

396. Groessl, M., S. Graf, and R. Knochenmuss, High resolution ion mobility-mass spectrometry for separation and identification of isomeric lipids. Analyst, 2015. **140**(20): p. 6904-6911.

397. O'Donnell, V.B., et al., LIPID MAPS: Serving the next generation of lipid researchers with tools, resources, data, and training. Science Signaling, 2019. **12**(563): p. eaaw2964.

398. Wojcik, R., et al., Lipid and Glycolipid Isomer Analyses Using Ultra-High Resolution Ion Mobility Spectrometry Separations. International journal of molecular sciences, 2017. **18**(1): p. 183.

399. Meier, F., et al., Parallel Accumulation–Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. Journal of Proteome Research, 2015. **14**(12): p. 5378-5387.

400. Kind, T., et al., LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat Methods, 2013. **10**.

401. Kochen, M.A., et al., Greazy: Open-Source Software for Automated Phospholipid Tandem Mass Spectrometry Identification. Analytical Chemistry, 2016. **88**(11): p. 5733-5741.

402. Kuhl, C., et al., CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. Analytical Chemistry, 2011. **84**(1): p. 283-289.

403. Käll, L., et al., Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. Journal of Proteome Research, 2007. **7**(1): p. 40-44.

404. Hsu, F.-F.u. and J. Turk, Characterization of ceramides by low energy collisional-activated dissociation tandem mass spectrometry with negative-ion electrospray ionization. Journal of the American Society for Mass Spectrometry, 2002. **13**(5): p. 558-570.

405. Ekroos, K., et al., Charting molecular composition of phosphatidylcholines by fatty acid scanning and ion trap MS3 fragmentation. Journal of Lipid Research, 2003. **44**(11): p. 2181-2192.

406. Hsu, F.-F. and J. Turk, Charge-driven fragmentation processes in diacyl glycerophosphatidic acids upon low-energy collisional activation. A mechanistic proposal. Journal of the American Society for Mass Spectrometry, 2000. **11**(9): p. 797-803.

407. Murphy, R.C., et al., Detection of the abundance of diacylglycerol and triacylglycerol molecular species in cells using neutral loss mass spectrometry. Analytical Biochemistry, 2007. **366**(1): p. 59-70.

408. Luo, P., et al., Multiple Reaction Monitoring-Ion Pair Finder: A Systematic Approach To Transform Nontargeted Mode to Pseudotargeted Mode for Metabolomics Study Based on Liquid Chromatography–Mass Spectrometry. Analytical Chemistry, 2015. **87**(10): p. 5050-5055.

409. Yan, Z. and R. Yan, Improved Data-Dependent Acquisition for Untargeted Metabolomics Using Gas-Phase Fractionation with Staggered Mass Range. Analytical Chemistry, 2015. **87**(5): p. 2861-2868.

410. Tsugawa, H., et al., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Meth, 2015. **12**(6): p. 523-526.

411. Gruner, T., et al., MOLGEN 4.0. MATCH Commun Math Comput Chem, 1998. **37**: p. 205-208.

412. Foster, J.M., et al., LipidHome: A Database of Theoretical Lipids Optimized for High Throughput Mass Spectrometry Lipidomics. PLoS ONE, 2013. **8**(5): p. e61951.

413. Hsu, F.-F. and J. Turk, Structural determination of sphingomyelin by tandem mass spectrometry with electrospray ionization. Journal of the American Society for Mass Spectrometry, 2000. **11**(5): p. 437-449.

414. Hsu, F.-F. and J. Turk, Characterization of ceramides by low energy collisional-activated dissociation tandem mass spectrometry with negative-ion electrospray ionization. Journal of the American Society for Mass Spectrometry, 2002. **13**(5): p. 558-570.

415. Kind, T., et al., LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat Meth, 2013. **10**(8): p. 755-758.

416. Kind, T., et al., LipidBlast Templates As Flexible Tools for Creating New in-Silico Tandem Mass Spectral Libraries. Analytical Chemistry, 2014. **86**(22): p. 11024-11027.

417. Ma, Y., et al., An in silico MS/MS library for automatic annotation of novel FAHFA lipids. Journal of cheminformatics, 2015. **7**: p. 53-53.

418. Stravs MA, Schymanski EL (2016) RMassBank Package. *http://www.bioconductor.org/packages/devel/bioc/html/RMassBank.html*. Accessed 14 Jan 2016.

419. Folch, J., M. Lees, and G.H.S. Stanley, A Simple Method For The Isolation And Purification Of Total Lipids From Animal Tissues. Journal of Biological Chemistry, 1957. **226**(1): p. 497-509.

420. Bodennec, J., et al., [9] - Purification of Sphingolipid Classes by Solid-Phase Extraction with Aminopropyl and Weak Cation Exchanger Cartridges, in Methods in Enzymology, A.H. Merrill and Y.A. Hannun, Editors. 2000, Academic Press. p. 101-114.

421. Shaner, R.L., et al., Quantitative analysis of sphingolipids for lipidomics using triple quadrupole and quadrupole linear ion trap mass spectrometers. Journal of lipid research, 2009. **50**(8): p. 1692-1707.

422. Ni, Z., et al., Computational solutions in redox lipidomics – Current strategies and future perspectives. Free Radical Biology and Medicine, 2019.

423. Wohlgemuth, G., et al., SPLASH, A hashed identifier for mass spectra. Nature biotechnology, 2016. **34**(11): p. 1099-1101.

424. Hsu, F.-F., Complete structural characterization of ceramides as [M−H]− ions by multiple-stage linear ion trap mass spectrometry. Biochimie, 2016. **130**: p. 63-75.

425. Chapter 6 Sphingolipids (SP), in Tandem Mass Spectrometry of Lipids: Molecular Analysis of Complex Lipids. 2015, The Royal Society of Chemistry. p. 194-232.

426. Schuhmann, K., et al., Quantitative Fragmentation Model for Bottom-up Shotgun Lipidomics. Analytical Chemistry, 2019.

427. Ni, Z., et al., LPPtiger software for lipidome-specific prediction and identification of oxidized phospholipids from LC-MS datasets. Scientific Reports, 2017. **7**(1): p. 15138.

428. Hernández-de-Diego, R., et al., PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Research, 2018. **46**(W1): p. W503-W509.

429. Gil de la Fuente, A., et al., Knowledge-based metabolite annotation tool: CEU Mass Mediator. Journal of Pharmaceutical and Biomedical Analysis, 2018. **154**: p. 138-149.

430. Gil-de-la-Fuente, A., et al., CEU Mass Mediator 3.0: A Metabolite Annotation Tool. Journal of Proteome Research, 2019. **18**(2): p. 797-802.

431. Shen, X., et al., Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nature Communications, 2019. **10**(1): p. 1516.

432. Naake, T. and E. Gaquerel, MetCirc: navigating mass spectral similarity in high-resolution MS/MS metabolomics data. Bioinformatics, 2017. **33**(15): p. 2419-2420.

433. Rainer, J., A. Vicini, and M. Witting. MetaboAnnotation. 2021; Available from: https://github.com/rformassspectrometry/MetaboAnnotation.

434. Rainer, J. and M. Witting. MetaboCoreUtils: Core Utils for Metabolomics Data. R package version 1.0.0. 2021; Available from: https://bioconductor.org/packages/release/bioc/html/MetaboCoreUtils.html.

435. Huang, N., et al., Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass

electrospray ionization mass spectral data. Journal of the American Society for Mass Spectrometry, 1999. **10**(11): p. 1166-1173.

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

437. Nothias, L.-F., et al., Feature-based molecular networking in the GNPS analysis environment. Nature Methods, 2020. **17**(9): p. 905-908.

438. Lerno, L.A., J.B. German, and C.B. Lebrilla, Method for the Identification of Lipid Classes Based on Referenced Kendrick Mass Analysis. Analytical Chemistry, 2010. **82**(10): p. 4236-4245.

439. Ashrafi, K., et al., Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes. Nature, 2003. **421**(6920): p. 268-272.

440. Mak, H.Y., et al., Polygenic control of Caenorhabditis elegans fat storage. Nat Genet, 2006. **38**(3): p. 363-368.

441. O'Rourke, E.J., et al., C. elegans Major Fats Are Stored in Vesicles Distinct from Lysosome-Related Organelles. Cell metabolism, 2009. **10**(5): p. 430-435.

442. Hu, F., et al., Live-cell vibrational imaging of choline metabolites by stimulated Raman scattering coupled with isotope-based metabolic labeling. Analyst, 2014. **139**(10): p. 2312-2317.

443. Wei, L., et al., Live-cell imaging of alkyne-tagged small biomolecules by stimulated Raman scattering. Nature Methods, 2014. **11**: p. 410.

444. Lange, M., et al., AdipoAtlas: A reference lipidome for human white adipose tissue. Cell Reports Medicine, 2021. **2**(10).

445. Symons, J.L., et al., Lipidomic atlas of mammalian cell membranes reveals hierarchical variation induced by culture conditions, subcellular membranes, and cell lineages. Soft Matter, 2021. **17**(2): p. 288-297.

446. Löfgren, L., et al., The BUME method: a novel automated chloroform-free 96-well total lipid extraction method for blood plasma. Journal of Lipid Research, 2012. **53**(8): p. 1690-1700.

447. Savini, M., et al., Lysosome lipid signalling from the periphery to neurons regulates longevity. Nature Cell Biology, 2022. **24**(6): p. 906-916.

448. Molenaars, M., et al., Metabolomics and lipidomics in Caenorhabditis elegans using a single-sample preparation. Disease Models & Mechanisms, 2021. **14**(4).

449. Spanier, B., et al., Comparison of lipidome profiles of Caenorhabditis elegans—results from an inter-laboratory ring trial. Metabolomics, 2021. **17**(3): p. 25.

450. Showalter, M.R., et al., The Emerging and Diverse Roles of Bis(monoacylglycero) Phosphate Lipids in Cellular Physiology and Disease. International Journal of Molecular Sciences, 2020. **21**(21): p. 8067.

451. Unfried, M., et al., LipidClock: A Lipid-Based Predictor of Biological Age. Frontiers in Aging, 2022. **3**.

452. Boland, S., et al., Phosphorylated glycosphingolipids essential for cholesterol mobilization in Caenorhabditis elegans. Nature Chemical Biology, 2017. **13**: p. 647.

453. Maceyka, M., et al., Sphingosine-1-phosphate signaling and its role in disease. Trends in Cell Biology, 2012. **22**(1): p. 50-60.

454. Helf, M.J., et al., Comparative metabolomics with Metaboseek reveals functions of a conserved fat metabolism pathway in C. elegans. Nature Communications, 2022. **13**(1): p. 782.

455. Yore, Mark M., et al., Discovery of a Class of Endogenous Mammalian Lipids with Anti-Diabetic and Anti-inflammatory Effects. Cell, 2014. **159**(2): p. 318-332.

456. Ma, Y., et al., An in silico MS/MS library for automatic annotation of novel FAHFA lipids. Journal of Cheminformatics, 2015. **7**(1): p. 53.

457. Nelson, A.B., et al., Artifactual FA dimers mimic FAHFA signals in untargeted metabolomics pipelines. Journal of Lipid Research, 2022. **63**(5): p. 100201.

458. Zhang, T., et al., A LC-MS–based workflow for measurement of branched fatty acid esters of hydroxy fatty acids. Nature Protocols, 2016. **11**(4): p. 747-763.

459. Benador, I.Y., et al., Mitochondria Bound to Lipid Droplets Have Unique Bioenergetics, Composition, and Dynamics that Support Lipid Droplet Expansion. Cell Metabolism, 2018. **27**(4): p. 869-885.e6.

460. Bowden, J.A., et al., Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using SRM 1950–Metabolites in Frozen Human Plasma. Journal of Lipid Research, 2017. **58**(12): p. 2275-2288.

461. Triebl, A., et al., Shared reference materials harmonize lipidomics across MS-based detection platforms and laboratories. Journal of Lipid Research, 2020. **61**(1): p. 105-115.

462. Gouveia, G.J., et al., Long-Term Metabolomics Reference Material. Analytical Chemistry, 2021. **93**(26): p. 9193-9199.

463. Klavins, K., et al., Quantitative Metabolite Profiling Utilizing Parallel Column Analysis for Simultaneous Reversed-Phase and Hydrophilic Interaction Liquid Chromatography Separations Combined with Tandem Mass Spectrometry. Analytical Chemistry, 2014. **86**(9): p. 4145-4150.

464. Schwaiger, M., et al., Merging metabolomics and lipidomics into one analytical run. Analyst, 2019. **144**(1): p. 220-229.

465. Chalcraft, K.R. and B.E. McCarry, Tandem LC columns for the simultaneous retention of polar and nonpolar molecules in comprehensive metabolomics analysis. Journal of Separation Science, 2013. **36**(21-22): p. 3478-3485.

466. Wang, Y., et al., Metabonomics Study on the Effects of the Ginsenoside Rg3 in a β-Cyclodextrin-Based Formulation on Tumor-Bearing Rats by a Fully Automatic Hydrophilic Interaction/Reversed-Phase Column-Switching HPLC−ESI-MS Approach. Analytical Chemistry, 2008. **80**(12): p. 4680-4688.

467. Greco, G., S. Grosse, and T. Letzel, Serial coupling of reversed-phase and zwitterionic hydrophilic interaction LC/MS for the analysis of polar and nonpolar phenols in wine. Journal of Separation Science, 2013. **36**(8): p. 1379-1388.

468. Greco, G., S. Grosse, and T. Letzel, Robustness of a method based on the serial coupling of reversed-phase and zwitterionic hydrophilic interaction LC–MS for the analysis of phenols. Journal of Separation Science, 2014. **37**(6): p. 630-634.

469. Haggarty, J., et al., Serially coupling hydrophobic interaction and reversed-phase chromatography with simultaneous gradients provides greater coverage of the metabolome. Metabolomics, 2015. **11**(5): p. 1465-1470.

470. Pyke, J.S., et al., A tandem liquid chromatography–mass spectrometry (LC–MS) method for profiling small molecules in complex samples. Metabolomics, 2015. **11**(6): p. 1552-1562.

471. Lv, W., et al., Pseudotargeted Method Based on Parallel Column Two-Dimensional Liquid Chromatography-Mass Spectrometry for Broad Coverage of Metabolome and Lipidome. Analytical Chemistry, 2020. **92**(8): p. 6043-6050.

472. Scholz, J., et al., Profiling of sphingolipids in Caenorhabditis elegans by two-dimensional multiple heart-cut liquid chromatography – mass spectrometry. Journal of Chromatography A, 2021. **1655**: p. 462481.

473. Willmann, L., et al., Metabolome analysis via comprehensive two-dimensional liquid chromatography: identification of modified nucleosides from RNA metabolism. Analytical and Bioanalytical Chemistry, 2015. **407**(13): p. 3555-3566.

474. Mohimani, H., et al., Dereplication of microbial metabolites through database search of mass spectra. Nature Communications, 2018. **9**(1): p. 4035.

475. Shannon, P., et al., Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research, 2003. **13**(11): p. 2498-2504.

476. Daniel, H., et al., Tandem HILIC-RP liquid chromatography for increased polarity coverage in food analysis. ELECTROPHORESIS, 2018. **0**(0).

477. Fuchs, S., et al., A metabolic signature of long life in Caenorhabditis elegans. BMC Biol, 2010. **8**.

478. Artyukhin, A.B., et al., Starvation-induced collective behavior in C. elegans. Scientific Reports, 2015. **5**(1): p. 10647.

479. Han, S., et al., Mono-unsaturated fatty acids link H3K4me3 modifiers to C. elegans lifespan. Nature, 2017. **544**(7649): p. 185-190.

480. Kleiman, R., G.F. Spencer, and F.R. Earle, Boron trifluoride as catalyst to prepare methyl esters from oils containing unusual acyl groups. Lipids, 1969. **4**(2): p. 118-122.

481. Hellmuth, C., et al., Nonesterified Fatty Acid Determination for Functional Lipidomics: Comprehensive Ultrahigh Performance Liquid Chromatography–Tandem Mass Spectrometry Quantitation, Qualification, and Parameter Prediction. Analytical Chemistry, 2012. **84**(3): p. 1483-1490.

482. B. Gowda, S.G., et al., Chemical Labeling Assisted Detection and Identification of Short Chain Fatty Acid Esters of Hydroxy Fatty Acid in Rat Colon and Cecum Contents. Metabolites, 2020. **10**(10): p. 398.

483. Ding, J., et al., In-Silico-Generated Library for Sensitive Detection of 2-Dimethylaminoethylamine Derivatized FAHFA Lipids Using High-Resolution Tandem Mass Spectrometry. Analytical Chemistry, 2020. **92**(8): p. 5960-5968.

484. Bollinger, J.G., et al., Improved Sensitivity Mass Spectrometric Detection of Eicosanoids by Charge Reversal Derivatization. Analytical Chemistry, 2010. **82**(16): p. 6790-6796.

485. Bollinger, J.G., et al., Liquid Chromatography/Electrospray Mass Spectrometric Detection of Fatty Acid by Charge Reversal Derivatization with More Than 4-Orders of Magnitude Improvement in Sensitivity. Journal of Lipid Research, 2013.

486. Wang, M., R.H. Han, and X. Han, Fatty Acidomics: Global Analysis of Lipid Species Containing a Carboxyl Group with a Charge-Remote Fragmentation-Assisted Approach. Analytical Chemistry, 2013. **85**(19): p. 9312-9320.

487. Li, Y., et al., C. elegans Fatty Acid Two-Hydroxylase Regulates Intestinal Homeostasis by Affecting Heptadecenoic Acid Production. Cellular Physiology and Biochemistry, 2018. **49**(3): p. 947-960.

488. Kespohl, M., et al., The Microbial Metabolite Butyrate Induces Expression of Th1-Associated Factors in CD4+ T Cells. Frontiers in Immunology, 2017. **8**.

489. Nazyrova, A., et al., HMDB: a knowledgebase for the human metabolome. Nucleic Acids Research, 2008. **37**(suppl_1): p. D603-D610.

490. Hansen, H.S., et al., Formation ofN-acyl-phosphatidylethanolamine andN-acylethanolamine (including anandamide) during glutamate-induced neurotoxicity. Lipids, 1999. **34**(S1Part3): p. S327-S330.

491. Triebl, A., et al., Quantitative analysis of N-acylphosphatidylethanolamine molecular species in rat brain using solid-phase extraction combined with reversed-phase

chromatography and tandem mass spectrometry. Journal of Separation Science, 2016. **39**(13): p. 2474-2480.

492. Yilmaz, L.S., et al., Modeling tissue-relevant Caenorhabditis elegans metabolism at network, pathway, reaction, and metabolite levels. Molecular Systems Biology, 2020. **16**(10): p. e9649.

493. Dirksen, P., et al., CeMbio - The Caenorhabditis elegans Microbiome Resource. G3 Genes|Genomes|Genetics, 2020. **10**(9): p. 3025-3039.

494. King, A.M., et al., Development of a rapid profiling method for the analysis of polar analytes in urine using HILIC–MS and ion mobility enabled HILIC–MS. Metabolomics, 2019. **15**(2): p. 17.

495. Munjoma, N., et al., High Throughput LC-MS Platform for Large Scale Screening of Bioactive Polar Lipids in Human Plasma and Serum. Journal of Proteome Research, 2022. **21**(11): p. 2596-2608.

496. Tanna, N., et al., Enhanced chromatographic efficiency obtained with vacuum jacketed columns facilitates the rapid UHPLC/MS/MS-based analysis of fasiglifam in rat plasma. Talanta, 2023. **254**: p. 124089.

497. Lu, Y., Z. Pang, and J. Xia, Comprehensive investigation of pathway enrichment methods for functional interpretation of LC–MS global metabolomics data. Briefings in Bioinformatics, 2022. **24**(1).

498. Rutz, A., et al., The LOTUS initiative for open knowledge management in natural products research. eLife, 2022. **11**: p. e70780.

499. Gray, N., et al., Development of a Rapid Microbore Metabolic Profiling Ultraperformance Liquid Chromatography–Mass Spectrometry Approach for High-Throughput Phenotyping Studies. Analytical Chemistry, 2016. **88**(11): p. 5742-5751.

500. Plumb, R.S., et al., High Throughput UHPLC-MS-Based Lipidomics Using Vacuum Jacketed Columns. Journal of Proteome Research, 2022. **21**(3): p. 691-701.

501. King, A.M., et al., Rapid profiling method for the analysis of lipids in human plasma using ion mobility enabled-reversed phase-ultra high performance liquid chromatography/mass spectrometry. Journal of Chromatography A, 2020. **1611**: p. 460597.

502. Danne-Rasche, N., C. Coman, and R. Ahrends, Nano-LC/NSI MS Refines Lipidomics by Enhancing Lipid Coverage, Measurement Sensitivity, and Linear Dynamic Range. Analytical Chemistry, 2018. **90**(13): p. 8093-8101.

503. Vasilopoulou, C.G., et al., Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts. Nature Communications, 2020. **11**(1): p. 331.

504. Alka, O., et al., DIAMetAlyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics. Nature Communications, 2022. **13**(1): p. 1347.

505. Meier, F., et al., diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. Nature Methods, 2020. **17**(12): p. 1229-1236.

506. Skowronek, P., et al., Synchro-PASEF allows precursor-specific fragment ion extraction and interference removal in data-independent acquisition. bioRxiv, 2022: p. 2022.11.01.514654.

507. Distler, U., et al., midiaPASEF maximizes information content in data-independent acquisition proteomics. bioRxiv, 2023: p. 2023.01.30.526204.

508. Commisso, M., et al., Performance comparison of electrospray ionization and atmospheric pressure chemical ionization in untargeted and targeted liquid chromatography/mass spectrometry based metabolomics analysis of grapeberry metabolites. Rapid Communications in Mass Spectrometry, 2017. **31**(3): p. 292-300.

509. Nordström, A., et al., Multiple Ionization Mass Spectrometry Strategy Used To Reveal the Complexity of Metabolomics. Analytical Chemistry, 2008. **80**(2): p. 421-429.

510. Campbell, J.L. and T. Baba, Near-Complete Structural Characterization of Phosphatidylcholines Using Electron Impact Excitation of Ions from Organics. Analytical Chemistry, 2015. **87**(11): p. 5837-5845.

511. Krumsiek, J., et al., Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Systems Biology, 2011. **5**(1): p. 21.

8. Appendix

8.1. Academic CV

8.1.1. CV

8.1.1.1. Personal Information

Name:	Dr. rer. nat. Witting, Michael Anton
Nationality:	German
Day and Place of Birth:	06.05.1986, Garmisch-Partenkirchen, Germany
Family Status:	married

8.1.1.2. Current Positions

Since 01/2021	Co-Head Metabolomics and Proteomics Core, Scientific Lead Metabolomics, Metabolomics and Proteomics Core, Helmholtz Zentrum München - German Research Center Environmental Health, Neuherberg, Germany
since 11/2015:	“Habilitand” at the Chair of Analytical Food Chemistry, TUM School of Life Sciences, Technical University of Munich, Freising-Weihenstephan, Germany
since 01/2016:	Freelance Lecturer for LifeScience-Akademie Bichlmeier, Munich, Germany
since 05/2018:	Training officer for biology and chemistry technician trainees Helmholtz Zentrum München - German Research Center Environmental Health, Neuherberg, Germany

8.1.1.3. Previous Positions

11/2015 – 12/2020:	Scientist at the Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München - German Research Center Environmental Health, Neuherberg, Germany
12/2012 – 10/2015:	Postdoctoral fellow at the Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München - German Research Center Environmental Health, Neuherberg, Germany
12/2009 – 11/2012:	PhD student at the Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München - German Research Center Environmental Health, Neuherberg, Germany
12/2009 – 11/2012:	Scientific Assistant at the Chair of genome-oriented Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising-Weihenstephan, Germany

8.1.1.4. Education

12/2009 – 11/2012:	Doctoral thesis at the Chair of Analytical Food Chemistry, TUM School of Life Sciences, Technical University of Munich, Freising-Weihenstephan, Germany, topic: “Host-pathogen metabolomics of Pseudomonas aeruginosa infection models”, grade: summa cum laude
05/2009 – 10/2009:	Diploma thesis at the Laboratory for Bioanalytical Chemistry, Georg-Simon-Ohm University of Applied Sciences, Nuremberg Germany, topic: “Bioanalytischer Nachweis von Histamin in biologischen Matries” (engl. Bioanalytical Detection of Histamine in Biological Matrices), grade: 1.0
10/2005 – 10/2009:	Studies of Applied Chemistry with functional direction Biochemistry, Georg-Simon-Ohm University of Applied Sciences, Nuremberg Germany, Final grade: 1.4

8.1.1.5. Awards and Prizes

10/2018:	Powerlist 2018: Top 40 under 40 young scientists in analytical science, https://theanalyticalscientist.com/power-list/2018
12/2017 – 11/2019:	Add-On Fellowship for Interdisciplinary Science der Joachim Herz Stiftung
04/2017:	Finalist Bioanalysis New Investigator Award 2017
01/2016 – 12/2018:	Postdoctoral Fellowship Program II (2016-2018) for highly talented early career scientists (HMGU PFP), Helmholtz Zentrum München, German Research Center Environmental Health, Neuherberg, Germany

8.1.1.6. Professional Involvement

Since 10/2020:	Member Board of Directors, Metabolomics Society
10/2023:	Organization of Munich Metabolomics Meeting 2022 with Prof. Strittmatter, Dr. Karin Kleigrewe, Dr. Jeannie Horak
10/2022:	Organization of Munich Metabolomics Meeting 2022 with Prof. Strittmatter, Dr. Karin Kleigrewe, Dr. Jeannie Horak
11/2020:	Organization of 7th Munich Metabolomics Meeting with Prof. Dr. Jerzy Adamski, Dr. Karin Kleigrewe and biocrates ag
05/2020	Member of Scientific Advisory Board Non-Target Analysis for Comprehensive Environmental Assessment 2020, Durham, USA (postponed to 05/2021 due to COVID-19 Pandemic)

8.1.1.7. Peer Review and Editorial Work

since 06/2022	Research Topic Editor in collaboration with Dr. Evelyn Rampler, “Reviews in Metabolomics: The Power of Separation Science in Metabolomics and Lipidomics”
04/2020 – 04/2021	Special Issue Guest Editor in collaboration with Prof. Horst-Joachim Schirra, Caenorhabditis elegans Applied to Metabolism Research, MDPI Metabolites
11/2017 – 06/2018:	Research Topic Editor in collaboration with Prof. Dr. Frank Schroeder and Prof. Dr. Arthur Edison, “Caenorhabditis elegans – Developing a Model Organism for Metabolomics, Lipidomics, and Glycomics”, Frontiers in Molecular Biology
01/2016 – 12/2017:	Special Issue Guest Editor in collaboration with Dr. Christophe Junot, “Identification of molecules from non-targeted analysis”, Journal of Chromatography B
since 01/2013:	Reviewer for different journals, e.g. Nature, Nature Communications, Science Advances, Journal of Chromatography A, Journal of Chromatography B, Analytical Chemistry, Electrophoresis, Metabolites, etc.

8.1.1.8. Professional Memberships

since 2021:	Deutsche Gesellschaft für Metabolomeforschung, Munich, Germany
since 2015:	Deutsche Gesellschaft für Massenspektrometrie, incl. „Fachgruppe LC-MS“, Bremen, Germany
since 2013:	Metabolomics Society, incl. Computational Mass Spectrometry Task Group, USA
since 2010:	Gesellschaft deutscher Chemiker (GdCh), incl. „Fachgruppe Analytische Chemie“ and „Arbeitskreis Separation Science“, Frankfurt am Main

8.1.1.9. Third Party Funding

Since 03/2023	HT-ADVANCE – Improving treatment efficacy in hypertension by biomarker-guided personalised decision support (HORIZON-HLTH-2022-TOOL-11) Total Budget: 13,477,342.00 € Own Budget: 547,315.00 €.
Since 02/2023	HUMAN – Harmonising and Unifying Blood Metabolomics Analysis Networks (HORIZON-MSCA-2021-DN-01-01) Total Budget: 2,582,503.20 € Own Budget: 521,078.40 €

Since 11/2022	Biomic_AUTh, Center of Excellence in Metabolomics research (HORIZON-WIDERA-2021-ACCESS-03) Total Budget: 1,226,702.50 € Own Budget: 121,510.00 €
01/2020 – 12/2022	Transferable retention time prediction for Liquid Chromatography-Mass Spectrometry-based metabolomics (DFG Research Grant) Total Budget: 401,090.00 € Own Budget: 218,550.00 €
03/2020 – 02/2023	MetClassNet: new approaches to bridge the gap between genome-scale metabolic networks and untargeted metabolomics (DFG Research Grant) Total Budget: 928,301.83 € Own Budget: 237,332.70 €

8.1.1.10. Project Management and Leadership Skills

09/2017:	Networking and Negotiating Skills (organized by HMGU PFP)
11/2016:	Time and Self-Management and Work Organization (organized by HMGU PFP)
10/2016:	Developing a career and personal research strategy (organized by HMGU PFP)
07/2016:	Team and Conflict Management (organized by HMGU PFP)
01/2016:	Bilateral Leadership (organized by HMGU PFP)

8.1.1.11. Didactic Education

09/2016 – 05/2018:	“ProLehre Intensiv 2016” (organized by TUM ProLehre)
07/2016:	“Grundlagen professioneller Hochschullehre” (organized by TUM ProLehre)
07/2016:	“Student-Centered Learning als Antwort auf die Heterogenität Studierender?” (organized by LMU Profil)
06/2016:	“Lehren im Labor” (organized by TUM ProLehre)
04/2016:	“Rhetorik und Präsentation” (organized by LMU Profil)

8.1.2. Publication list

8.1.2.1. Under Review / Under Revision / Preprints

SIN-3 transcriptional coregulator maintains mitochondrial hormesis and polyamine flux

Giovannetti M., P. Fabrizio, O. Nicolle, C. Bedet, M. Rodríguez-Palero, G. Michaux, M. Artal-Sanz, M. Witting, F. Palladino

J Cell Biol, under review

RepoRT: A comprehensive repository for small molecule retention times

Kretschmer F., E.-M. Harrieder, M. Hoffmann, S. Böcker, M. Witting

ChemRxiv, https://chemrxiv.org/engage/chemrxiv/article-details/64a5a08c9ea64cc1677e120f

Nature Methods, under review

APEX – an annotation propagation workflow through multiple experimental networks to improve the annotation of new metabolite classes in Caenorhabditis elegans

Salzer L., E. M. Novoa-del-Toro, C. Frainay, K. A. B. Kissoyan, F. Jourdan, K. Dierking, M. Witting

Anal Chem, under review

(Re-)use and (re-)analysis of publicly available metabolomics data

Witting M.

Proteomics, under review

(Metabolomic analysis of the mouse host reveals glutamyl-glutamate as mediator of pro-reproductive effects of Lactiplantibacillus plantarum SNI3

Sun N., B. Juhász, K. Horváth, D. Kuti, I. Bata-Vidács, J. Shen, A. Feuchtinger, I. Nagy, J. Kukolya, S. Ferenczi, M. Witting, A. Walch, K. J. Kovács

elife, under review

8.1.2.2. 2023

Capillary Electrophoresis - Mass Spectrometry as tool for Caenorhabditis elegans metabolomics research

Salzer L., P. Schmitt-Kopplin, M. Witting

Metabolomics. 2023 Jun 23;19(7):61

Lipidomic and Metallomic Alteration of Caenorhabditis elegans after Acute and Chronic Manganese, Iron, and Zinc Exposure with a Link to Neurodegenerative Disorders

Blume B., V. Schwantes, M. Witting, H. Hayen, P. Schmitt-Kopplin, P. O. Helmer, B. Michalke

J Proteome Res. 2023 Mar 3;22(3):837-850.

8.1.2.3. 2022

Providing metabolomics education and training: pedagogy and considerations

Winder C. L., M. Witting, F. Tugizimana, W. Dunn and S. N. Reinke, on behalf of the Metabolomics Society Education and Training Committee

Metabolomics. 2022 Dec 13;18(12):106.

Critical assessment of chromatographic metadata in publicly available metabolomics data repositories

Harrieder E.-M., F. Kretschmer, W. Dunn, S. Böcker, M. Witting

Metabolomics. 2022 Nov 27;18(12):97.

Quality Assurance and Quality Control Reporting in Untargeted Metabolic Phenotyping: mQACC Recommendations for Analytical Quality Management

Kirwan J., H. Gika, R. Beger, D. Bearden, W. Dunn, R. Goodacre, G. Theodoridis, M. Witting, L.-R. Yu, I. Wilson

Metabolomics. 2022 Aug 27;18(9):70.

Introducing the Lipidomics Minimal Reporting Checklist.

McDonald J., C Ejsing, D. Kopczynski, M. Holcapek, R. Ahrends, J. Aoki, M. Arita, M. Arita, E. Baker, J. Bertrand-Michel, J. Bowden, B. Brügger, S. Ellis, M. Fedorova, W. Griffiths, X. Han, J. Hartler, N. Hoffmann, J. Koelmel, H. Koefeler, T. Mitchell, V. O'Donnell, D. Saigusa, D. Schwudke, A. Shevchenko, C. Ulmer, M. Wenk, M. Witting, D. Wolrab, Y. Xia, G. Liebisch

Nat Metab. 2022 Sep;4(9):1086-1088.

MobilityTransformR: An R package for effective mobility transformation of CE-MS data

Salzer L., M. Witting, P. Schmitt-Kopplin

Bioinformatics. 2022 Jul 4;btac441

Networks and graph discovery in metabolomics data analysis and interpretation

Amara A., C. Frainay. F. Jourdan T. Naake, S. Neumann, E. M. Novoa-del-Toro, R. Salek, L. Salzer, S. Scharfenberg, M. Witting

Front Mol Biosci. 2022 Mar 8;9:841373.

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

Alka O., P. Shanthamoorthy, M. Witting, K. Kleigrewe, O. Kohlbacher, H. L. Röst

Nat Commun. 2022 Mar 15;13(1):1347

A modular and expandable ecosystem for metabolomics data annotation in R

Rainer J., A. Vicini, L. Salzer, J. Stanstrup, J. M. Badia, S. Neumann, M. Stravs, V. V. Hernandes, L. Gatto, S. Gibb, M. Witting

Metabolites. 2022 Feb 11;12(2):173.

Impaired phosphocreatine metabolism in white adipocytes promotes inflammation

Maqdasy S., S. Lecoutre, G. Renzi, S. Frendo-Cumbo, D. Rizo-Roca, T. Moritz, M. Juvany, O. Hodek, H. Gao, M. Couchet, M. Witting, A. Kerr, M. O. Bergo, R. P. Choudhury, M. Aouadi, J. R. Zierath, A. Krook, N. Mejhert, M. Rydén

Nat Metab. 2022 Feb;4(2):190-202

8.1.2.4. 2021

Novel extraction method for combined lipid and metal speciation from Caenorhabditis elegans with focus on iron redox status and lipid profiling

Blume B., M. Witting, P. Schmitt-Kopplin, B. Michalke

Front Chem. 2021 Dec 9;9:788094.

N-Alkylpyridinium sulfonates for retention time indexing in reversed-phase-liquid chromatography-mass spectrometry-based metabolomics

Stoffel R., M. Quilliam, N. Hardt, A. Fridstrom, M. Witting

Anal Bioanal Chem. 2021 Dec 15

Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics

Harrieder E.-M., F. Kretschmer, S. Böcker, M. Witting

J Chromatogr B Analyt Technol Biomed Life Sci. 2022 Jan 1;1188:123069.

High-confidence structural annotation of metabolites absent from spectral libraries

Hoffmann M. A, L. F. Nothias, M. Ludwig, M. Fleischauer, E. C. Gentry, M. Witting, P. C. Dorrestein, K. Dührkop, S. Böcker

Nat Biotechnol. 2021 Oct 14. doi: 10.1038/s41587-021-01045-9.

Quo Vadis Caenorhabditis elegans Metabolomics-A Review of Current Methods and Applications to Explore Metabolism in the Nematode

Salzer L., M. Witting

Metabolites, 2021 Apr 29;11(5):284

HLH-30 dependent rewiring of metabolism during starvation in C. elegans

Dall K. B., J. F. Havelund, E. B. Harvald, M. Witting, N. J. Færgeman

Aging Cell, 2021 Apr;20(4):e13342

Comparison of lipidome profiles of Caenorhabditis elegans – Results from an inter-laboratory ring trial

Spanier B., A. Laurençon, A. Weiser, N. Pujol, S. Omi, A. Barsch, S. W. Meyer, J. J. Ewbank, F. Paladino, S. Garvis, H. Aguilaniu, M. Witting

Metabolomics, 2021 Feb 17;17(3):25.

UHPLC-IMS-Q-ToF-MS analysis of Maradolipids, found exclusively in Caenorhabditis elegans dauer larvae

Witting M., U. Schmidt, H.-J. Knölker

Analytical and Bioanalytical Chemistry, 2021 Mar;413(8):2091-2102.

IL-17 controls central nervous system autoimmunity through the intestinal microbiome

Regen T., S. Isaac, A. Amorim, N. G. Núñez, J. Hauptmann, A. Shanmugavadivu, M. Klein, R. Sankowski, I. A. Mufazalov, N. Yogev, J. Huppert, F. Wanke, M. Witting, A. Grill, E. J. C. Gálvez, A. Nikolaev, M. Blanfeld, I. Prinz, P. Schmitt-Kopplin, T. Strowig, C. Reinhardt, M. Prinz, T. Bopp, B. Becher, C. Ubeda, A. Waisman

Science Immunology, 2021 Feb 5;6(56):eaaz6563

Reduced peroxisomal import triggers a peroxisomal retrograde signaling

Rackles E., I. Forné, C. Fischer, X. Zhang, S. Schrott, J. Zacherl, M. Witting, J. Ewbank, C. Osman, A. Imhof, S. G. Rolland

Cell Reports, 2021 Jan 19;34(3):108653.

8.1.2.5. 2020

Comprehensive vitamer profiling of folate mono- and polyglutamates in baker's yeast (Saccharomyces cerevisiae) as a function of different sample preparation procedures

Gmelch L., D. Wirtz, M. Witting, N. Weber, L. Striegel, P. Schmitt-Kopplin, M. Rychlik

Metabolites, 2020 Jul 23;10(8):E301.

Metabolomic adjustments in the orchid mycorrhizal fungus Tulasnella calospora during symbiosis with Serapias vomeracea

Ghirardo A., V. Fochi, B. Lange, M. Witting, J.-P. Schnitzler, S. Perotto, R. Balestrini

New Phytologist, 2020 Jul 15. doi: 10.1111/nph.16812

Feature-based Molecular Networking in the GNPS Analysis Environment

Nothias L. F., D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa, M. Fleischauer, F. Aicheler, A. Aksenov, O. Alka, P.-M. Allard, A. Barsch, X. Cachet, M. Caraballo, R. R. Da Silva, T. Dang, N. Garg, J. M. Gauglitz, A. Gurevich, G. Isaac, A. K. Jarmusch, Z. Kameník, K. B. Kang, N. Kessler, I. Koester, A. Korf, A. Le Gouellec, M. Ludwig, M. H. Christian, L.-I. McCall, J. McSayles, S. W. Meyer, H. Mohimani, M. Morsy, O. Moyne, S. Neumann, H. Neuweger, N. H. Nguyen, M. Nothias-Esposito, J. Paolini, V. V. Phelan, T. Pluskal, R. A. Quinn, S. Rogers, B. Shrestha, A. T., J. J. J. van der Hooft, F. Vargas, K. C. Weldon, M. Witting, H. Yang, Z. Zhang, F. Zubeil, O. Kohlbacher, S. Böcker, T. Alexandrov, N. Bandeira, M. Wang, P. C. Dorrestein

Nature Methods, 2020 Sep;17(9):905-908

Suggestions for Standardized Identifiers for Fatty Acyl Compounds in Genome Scale Metabolic Models and Their Application to the WormJam Caenorhabditis elegans Model

Witting M.

Metabolites. 2020 Mar 28;10(4):E130. doi: 10.3390/metabo10040130.

Current status of retention time prediction in metabolite identification

Witting M., S. Böcker

Journal of Separation Science, 2020 Mar 7.

Autophagy compensates for defects in mitochondrial dynamics

Haeussler S., F. Köhler, M. Witting, M. F. Premm, S. G. Rolland, C. Fischer, L. Chauve, O. Casanueva, B. Conradt

PLoS Genetics, 2020 Mar 19;16(3):e1008638.

In-vivo targeted tagging of RNA isolates cell specific transcriptional responses to environmental stimuli and identifies liver-to-adipose RNA transfer

Darr J., M. Lassi, A. Tomar, R. Gerlini, F. Scheid, MH de Angelis, M. Witting, R. Teperino

Cell Reports, 2020 Mar 3;30(9):3183-3194.e4.

8.1.2.6. 2019

Development and application of a HILIC UHPLC-MS method for polar fecal metabolome profiling

Sillner N., A. Walker, EM. Harrieder, P. Schmitt-Kopplin, M. Witting

J Chromatogr B Analyt Technol Biomed Life Sci. 2019 Mar 1;1109:142-148. doi: 10.1016/j.jchromb.2019.01.016..

The metaRbolomics Toolbox in Bioconductor and beyond

Stanstrup J., CD. Broeckling, R. Helmus, N. Hoffmann, E. Mathé, T. Naake, L. Nicolotti, K. Peters, J. Rainer, RM. Salek, T. Schulze, E. Schymanski, MA. Stravs, EA. Thévenot, H. Treutler, RJM. Weber, E. Willighagen, M. Witting, S. Neumann

Metabolites. 2019 Sep 23;9(10). pii: E200. doi: 10.3390/metabo9100200.

The sphingolipidome of the model organism Caenorhabditis elegans

Hänel, V., C. Pendleton, M. Witting

Chem Phys Lipids. 2019 Aug;222:15-22. doi: 10.1016/j.chemphyslip.2019.04.009.

8.1.2.7. 2018

Mycorrhiza-Triggered Transcriptomic and Metabolomic Networks Impinge on Herbivore Fitness

Kaling M., A. Schmidt, F. Moritz, M. Rosenkranz, M. Witting, K. Kasper, D. Janz, P. Schmitt-Kopplin, JP. Schnitzler, A. Polle

Plant Physiol. 2018 Apr;176(4):2639-2656. doi: 10.1104/pp.17.01810. Epub 2018 Feb 8.

Modeling Meets Metabolomics-The WormJam Consensus Model as Basis for Metabolic Studies in the Model Organism Caenorhabditis elegans

Witting M., J. Hastings, N. Rodriguez, CJ. Joshi, JPN. Hattwell, PR. Ebert, M. van Weeghel, AW. Gao, MJO. Wakelam, RH. Houtkooper, A. Mains, N. Le Novère, S. Sadykoff, F. Schroeder, NE. Lewis, HJ. Schirra, C. Kaleta, O. Casanueva

Front Mol Biosci. 2018 Nov 14;5:96. doi: 10.3389/fmolb.2018.00096.

Usage of FT-ICR-MS Metabolomics for Characterizing the Chemical Signatures of Barrel-Aged Whisky

Roullier-Gall C., J. Signoret, D. Hemmler, M. Witting, B. Kanawati, B. Schäfer, RD. Gougeon, P. Schmitt-Kopplin P

Front Chem. 2018 Feb 22;6:29. doi: 10.3389/fchem.2018.00029.

Metabotype variation in a field population of tansy plants influences aphid host selection.

Clancy MV., SE. Zytynska, F. Moritz, M. Witting, P. Schmitt-Kopplin, WW. Weisser, JP. Schnitzler

Plant Cell Environ. 2018 Dec;41(12):2791-2805. doi: 10.1111/pce.13407. Epub 2018 Aug 17

Pharmacometabolic response to pirfenidone in pulmonary fibrosis detected by MALDI-FTICR-MSI

Sun N., IE. Fernandez, M. Wie, M. Witting, M. Aichler, A. Feuchtinger, G. Burgstaller, SE. Verleden, P. Schmitt-Kopplin, O. Eickelberg, A. Walch

Eur Respir J. 2018 Sep 15;52(3). pii: 1702314. doi: 10.1183/13993003.02314-2017

Metformin impacts cecal bile acid profiles in mice

Sillner N., A. Walker, W. Koch, M. Witting, P. Schmitt-Kopplin

J Chromatogr B Analyt Technol Biomed Life Sci. 2018 Apr 15;1083:35-43. doi: 10.1016/j.jchromb.2018.02.029.

Tandem HILIC-RP liquid chromatography for increased polarity coverage in food analysis

Hemmler D., SS. Heinzmann, K. Wöhr, P. Schmitt-Kopplin, M. Witting

Electrophoresis. 2018 Jul;39(13):1645-1653. doi: 10.1002/elps.201800038.

8.1.2.8. 2017

Amniotic Fluid and Maternal Serum Metabolic Signatures in the Second Trimester Associated with Preterm Delivery

Virgiliou C., HG. Gika, M. Witting, AA. Bletsou, A. Athanasiadis, M. Zafrakas, NS. Thomaidis, N. Raikos, G. Makrydimas, GA. Theodoridis

J Proteome Res. 2017 Feb 3;16(2):898-910. doi: 10.1021/acs.jproteome.6b00845.

Metabolic Profile of Human Coelomic Fluid

Virgiliou C., L. Valianou, M. Witting, F. Moritz, C. Fotaki, P. Zoumpoulakis, AC. Chatziioannou, L. Lazaros, G. Makrydimas, K. Chatzimeletiou, N. Raikos, GA. Theodorids

Bioanalysis. 2017 Jan;9(1):37-51. doi: 10.4155/bio-2016-0223.

Identification of a High-Affinity Pyruvate Receptor in Escherichia coli

Behr S., I. Kristoficova, M. Witting, EJ. Breland, AR. Eberly, C. Sachs, P. Schmitt-Kopplin, M. Hadjifrangiskou, K. Jung

Sci Rep. 2017 May 3;7(1):1388. doi: 10.1038/s41598-017-01410-2

LipidFrag: Improving reliability of in silico fragmentation of lipids and application to the Caenorhabditis elegans lipidome

Witting M., C. Ruttkies, S. Neumann, P. Schmitt-Kopplin

PLoS One. 2017 Mar 9;12(3):e0172311. doi: 10.1371/journal.pone.0172311

Comparative analysis of LytS/LytTR-type histidine kinase/response regulator systems in γ-proteobacteria

Behr S., S. Brameyer, M. Witting, P. Schmitt-Kopplin, K. Jung

PLoS One. 2017 Aug 10;12(8):e0182993. doi: 10.1371/journal.pone.0182993

QSRR Modeling for Metabolite Standards Analyzed by Two Different Chromatographic Columns Using Multiple Linear Regression

Zisi C., I. Sampsonidis, S. Fasoula, K. Papachristos, M. Witting, HG. Gika, P. Nikitas, A. Pappa-Louisi

Metabolites. 2017 Feb 9;7(1). pii: E7. doi: 10.3390/metabo7010007.

8.1.2.9. 2016

Natural oxygenation of Champagne wine during ageing on lees: A metabolomics picture of hormesis

Roullier-Gall C., M. Witting, F. Moritz, RB. Gil, D. Goffette, M. Valade, P. Schmitt-Kopplin, RD. Gougeon

Food Chem. 2016 Jul 15;203:207-215. doi: 10.1016/j.foodchem.2016.02.043.

The Caenorhabditis elegans lipidome: A primer for lipid analysis in Caenorhabditis elegans

Witting M., P. Schmitt-Kopplin

Arch Biochem Biophys. 2016 Jan 1;589:27-37. doi: 10.1016/j.abb.2015.06.003, 3.559

The Role of Dafachronic Acid Signaling in Development and Longevity in Caenorhabditis elegans: Digging Deeper Using Cutting-Edge Analytical Chemistry

Aguilaniu H., P. Fabrizio, M. Witting

Front Endocrinol (Lausanne). 2016 Feb 11;7:12. doi: 10.3389/fendo.2016.00012.

8.1.2.10. 2015

DI-ICR-FT-MS-based high-throughput deep metabotyping: a case study of the Caenorhabditis elegans-Pseudomonas aeruginosa infection model

Witting M., M. Lucio, D. Tziotis, B. Wägele, K. Suhre, R. Voulhoux, S. Garvis, P. Schmitt-Kopplin

Anal Bioanal Chem. 2015 Feb;407(4):1059-73. doi: 10.1007/s00216-014-8331-5.

Integrating analytical resolutions in non-targeted wine metabolomics

Roullier-Gall C., M. Witting, D. Tziotis, A. Ruf, RD. Gougeon, P. Schmitt-Kopplin

Tetrahedron. 2015 May;71(20):2983-2990.

Computational analysis and ratiometric comparison approaches aimed to assist column selection in hydrophilic interaction liquid chromatography-tandem mass spectrometry targeted metabolomics

Sampsonidis I., M. Witting, W. Koch, C. Virgiliou, HG. Gika, P. Schmitt-Kopplin, GA. Theodoridis

J Chromatogr A. 2015 Aug 7;1406:145-55. doi: 10.1016/j.chroma.2015.06.008. Epub 2015 Jun 14.

Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting

Fellner L., S. Simon, C. Scherling, M. Witting, S. Schober, C. Polte, P. Schmitt-Kopplin, DA. Keim, S. Scherer, K. Neuhaus

BMC Evol Biol. 2015 Dec 18;15:283. doi: 10.1186/s12862-015-0558-z.

Chemical messages in 170-year-old champagne bottles from the Baltic Sea: Revealing tastes from the past

Jeandet P., SS. Heinzmann, C. Roullier-Gall, C. Cilindre, A. Aron, MA. Deville, F. Moritz, T. Karbowiak, D. Demarville, C. Brun, F. Moreau, B. Michalke, G. Liger-Belair, M. Witting, M. Lucio, D. Steyer, RD. Gougeon, P. Schmitt-Kopplin.

Proc Natl Acad Sci U S A. 2015 May 12;112(19):5893-8. doi: 10.1073/pnas.1500783112

Fast separation and quantification of steroid hormones Δ4- and Δ7-dafachronic acid in Caenorhabditis elegans

Witting M., HC. Rudloff, M. Thondamal, H. Aguilaniu, P. Schmitt-Kopplin

J Chromatogr B Analyt Technol Biomed Life Sci. 2015 Jan 26;978-979:118-21. doi: 10.1016/j.jchromb.2014.12.005.

8.1.2.11. 2014

Distinct signatures of host-microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet

Walker A., B. Pfitzner, S. Neschen, M. Kahle, M. Harir, M. Lucio, F. Moritz, D. Tziotis, M. Witting, M. Rothballer, M. Engel, M. Schmid, D. Endesfelder, M. Klingenspor, T. Rattei, WZ. Castell, MH. de Angelis, A. Hartmann, P. Schmitt-Kopplin

ISME J. 2014 Dec;8(12):2380-96. doi: 10.1038/ismej.2014.79.

Molecular and structural characterization of dissolved organic matter during and post cyanobacterial bloom in Taihu by combination of NMR spectroscopy and FTICR mass spectrometry

Zhang F., M. Harir, F. Moritz, J. Zhang, M. Witting, Y. Wu, P. Schmitt-Kopplin, A. Fekete, A. Gaspar, N. Hertkorn

Water Res. 2014 Jun 15;57:280-94. doi: 10.1016/j.watres.2014.02.051.

High-resolution metabolite imaging of light and dark treated retina using MALDI-FTICR mass spectrometry

Sun N., A. Ly, S. Meding, M. Witting, SM. Hauck, M. Ueffing, P. Schmitt-Kopplin, M. Aichler, A. Walch

Proteomics. 2014 Apr;14(7-8):913-23. doi: 10.1002/pmic.201300407.

Ultrahigh resolution mass spectrometry-based metabolic characterization reveals cerebellum as a disturbed region in two animal models

Lin S., B. Kanawati, L. Liu, M. Witting, M. Li, J. Huang, P. Schmitt-Kopplin, Z. Cai

Talanta. 2014 Jan;118:45-53. doi: 10.1016/j.talanta.2013.09.019. Epub 2013 Oct 5.

Phenotype of htgA (mbiA), a recently evolved orphan gene of Escherichia coli and Shigella, completely overlapping in antisense to yaw

Fellner L., N. Bechtel, M. Witting, S. Simon, P. Schmitt-Kopplin, DA. Keim, S. Scherer, K. Neuhaus

FEMS Microbiol Lett. 2014 Jan;350(1):57-64. doi: 10.1111/1574-6968.12288.

High precision mass measurements for wine metabolomics

Roullier-Gall C.*, M. Witting*, RD. Gougeon, P. Schmitt-Kopplin

Front Chem. 2014 Nov 13;2:102. doi: 10.3389/fchem.2014.00102.

* equally contributed

Steroid hormone signalling links reproduction to lifespan in dietary-restricted Caenorhabditis elegans

Thondamal M., M. Witting, P. Schmitt-Kopplin, H. Aguilaniu

Nat Commun. 2014 Sep 11;5:4879. doi: 10.1038/ncomms5879.

Optimizing a ultrahigh pressure liquid chromatography-time of flight-mass spectrometry approach using a novel sub-2μm core-shell particle for in depth lipidomic profiling of Caenorhabditis elegans

Witting M., TV. Maier, S. Garvis, P. Schmitt-Kopplin

J Chromatogr A. 2014 Sep 12;1359:91-9. doi: 10.1016/j.chroma.2014.07.02.

8.1.2.12. 2012

MassTRIX reloaded: combined analysis and visualization of transcriptome and metabolome data

Wägele B., M. Witting, P. Schmitt-Kopplin, K. Suhre

PLoS One. 2012;7(7):e39860. doi: 10.1371/journal.pone.0039860.

8.1.3. Book Chapter

8.1.3.1. 2020

Using Genome-Scale Metabolic Networks for Analysis, Visualization, and Integration of Targeted Metabolomics Data

Hattwell J. P.N., J. Hastings, O. Casanueva, H. J. Schirra, M. Witting

Methods Mol Biol. 2104; 361-386. doi: 10.1007/978-1-0716-0239-3_18

8.1.3.2. 2018

Bio- and Chemoinformatics Approaches for Metabolomics Data Analysis

Witting M.

Methods Mol Biol. 2018;1738:41-61. doi: 10.1007/978-1-4939-7643-0_4

8.1.3.3. 2015

Combined Nontargeted Analytical Methodologies for the Characterization of the Chemical Evolution of Bottled Wines

Roullier-Gall C., M. Witting, D. Tziotis, A. Ruf, M. Lucio, P. Schmitt-Kopplin, R. D. Gougeon

Advances in Wine Research, Chapter 2, 13-27

8.1.3.4. 2014

Transcriptome and Metabolome Data Integration – Technical Perquisites for Successful Data Fusion and Visualization

Witting M., P. Schmitt-Kopplin

Fundamentals of Advanced Omics Technologies: From Genes to Metabolites. C. Simo, A. Cifuentse, V. Garcia-Canas, Elsevier Heidelberg: 421-442.

8.1.3.5. 2012

Ultrahigh Resolution Mass Spectrometry Based Non-targeted Microbial Metabolomics

Witting M., M. Lucio, D. Tziotis and P. Schmitt-Kopplin

Genetics Meets Metabolomics. K. Suhre, Springer New York: 57-71.

8.1.4. Non-Peer Reviewed Articles

Identification of molecules from non-targeted analysis

Junot C., M. Witting

J Chromatogr B Analyt Technol Biomed Life Sci. 2017 Dec 15;1071:1-2. doi: 10.1016/j.jchromb.2017.

Probenvorbereitung für die LC-MS basierte Metabolomik und Lipidomik (engl. Sample preparation for LC-MS based metabolomics and lipidomics)

Witting M.

GIT-Labor – Portal für Anwender in Wissenschaft und Industrie

https://www.git-labor.de/forschung/chemie-physik/probenvorbereitung-fuer-die-lc-ms-basierte-metabolomik-und-lipidomik

Landmark Literature 2018: Part I – Phosphate to the Rescue

Witting M.

The Analytical Scientist

https://theanalyticalscientist.com/fields-applications/landmark-literature-2018

Investigating the increased lifespan in C. elegans daf-2 mutants by 4D-Lipidomics

Witting M., A. Barsch, S. W. Meyer, U. Schweiger-Hufnagel, N. Kessler, P. Schmitt-Kopplin

Bruker Application Note

Combination of stationary phase selectivity in SFC method development

Bieber S., P. Schmitt-Kopplin, M. Witting, T. Letzel

AFIN-TS Forum; April (3): 1-18.

8.2. Supplementary Tables

The following tables summarize all metabolites and lipids used in chapter 3 for the analysis of RT, RI and CCS.

8.2.1. Metabolites

SI Table 1: Metabolites contained in the MSMLS standard collection used for analysis of RT, RI and CCS throughout chapter 3.

Name	formula	exact mass	logP	Mix
β-Nicotinamide adenine dinucleotide	$C_{21}H_{27}N_7O_{14}P_2$	663.1091	-9.87	P1_A
Glutamine	$C_5H_{10}N_2O_3$	146.0691	-4.00	P1_A
Hypotaurine	$C_2H_7NO_2S$	109.0197	-2.70	P1_A
IMP	$C_{10}H_{13}N_4O_8P$	348.0471	-2.93	P1_A
Citric acid	$C_6H_8O_7$	192.0270	-1.32	P1_A
Threonine	$C_4H_9NO_3$	119.0582	-3.47	P1_A
Purine	$C_5H_4N_4$	120.0436	-0.34	P1_A
N-Acetylneuraminic acid	$C_{11}H_{19}NO_9$	309.1060	-3.56	P1_A
Kynurenine	$C_{10}H_{12}N_2O_3$	208.0848	-1.91	P1_A
Pyrimidine	$C_4H_4N_2$	80.0374	0.05	P1_A
Aspartic acid	$C_4H_7NO_4$	133.0375	-3.50	P1_A
Uric acid	$C_5H_4N_4O_3$	168.0283	-1.54	P1_A
Cytidne	$C_9H_{13}N_3O_5$	243.0855	-2.80	P1_B
Serine	$C_3H_7NO_3$	105.0426	-3.89	P1_B
Cysteine	$C_3H_7NO_2S$	121.0197	-2.79	P1_B
Citrulline	$C_6H_{13}N_3O_3$	175.0957	-3.93	P1_B
Taurine	$C_2H_7NO_3S$	125.0147	-2.61	P1_B
Gluconolactone	$C_6H_{10}O_6$	178.0477	-2.75	P1_B

Name	formula	exact mass	logP	Mix
Nicotinic acid	$C_6H_5NO_2$	123.0320	-0.17	P1_B
Inosine	$C_{10}H_{12}N_4O_5$	268.0808	-2.48	P1_B
γ-Aminobutyric acid	$C_4H_9NO_2$	103.0633	-2.89	P1_B
Cytosine	$C_4H_5N_3O$	111.0433	-1.15	P1_B
Isoleucine	$C_6H_{13}NO_2$	131.0946	-1.51	P1_B
Pyrazole	$C_3H_4N_2$	68.0374	0.28	P1_B
Glutamic acid	$C_5H_9NO_4$	147.0532	-3.24	P1_C
Ascorbic acid	$C_6H_8O_6$	176.0321	-1.91	P1_C
p-Hydroxyphenylacetic acid	$C_8H_8O_3$	152.0468	-3.17	P1_C
N-Acetylglucosamine	$C_8H_{15}NO_6$	221.0899	-3.22	P1_C
Glycolic acid	$C_2H_4O_3$	76.0160	-1.04	P1_C
Sarcosine	$C_3H_7NO_2$	89.0477	-3.19	P1_C
Gluconic acid	$C_6H_{12}O_7$	196.0583	-3.41	P1_C
Quinic acid	$C_7H_{12}O_6$	192.0634	-2.70	P1_C
Dihydroorotic acid	$C_5H_6N_2O_4$	158.0328	-1.52	P1_C
Malonic acid	$C_3H_4O_4$	104.0110	-0.33	P1_C
Pipecolic acid	$C_6H_{11}NO_2$	129.0790	-2.12	P1_C
Formamide	CH_3NO	45.0215	-1.08	P1_C
Glycine	$C_2H_5NO_2$	75.0320	-3.41	P1_D
Methionine	$C_5H_{11}NO_2S$	149.0510	-2.19	P1_D
Tetrahydrofolic acid	$C_{19}H_{23}N_7O_6$	445.1710	-2.22	P1_D
Adenine	$C_5H_5N_5$	135.0545	-0.57	P1_D
Mehtylthioadenosine	$C_{11}H_{15}N_5O_3S$	297.0896	-0.61	P1_D

Name	formula	exact mass	logP	Mix
Thymidine	$C_{10}H_{14}N_2O_5$	242.0903	-1.12	P1_D
Glyceric acid	$C_3H_6O_4$	106.0266	-1.52	P1_D
Orotic acid	$C_5H_4N_2O_4$	156.0171	-1.23	P1_D
Ethanolamine phosphate	$C_2H_8NO_4P$	141.0191	-2.45	P1_D
Xanthine	$C_5H_4N_4O_2$	152.0334	-0.21	P1_D
Dihydrofolic acid	$C_{19}H_{21}N_7O_6$	443.1553	-1.44	P1_D
Cystine	$C_6H_{12}N_2O_4S_2$	240.0238	-5.90	P1_D
Alanine	$C_3H_7NO_2$	89.0477	-2.84	P1_E
Tryptophan	$C_{11}H_{12}N_2O_2$	204.0899	-1.09	P1_E
UMP	$C_9H_{13}N_2O_9P$	324.0359	-2.54	P1_E
Proline	$C_5H_9NO_2$	115.0633	-2.57	P1_E
Thymine	$C_5H_6N_2O_2$	126.0429	-0.46	P1_E
Succinate Semialdehyde	$C_4H_6O_3$	102.0317	-0.56	P1_E
Lactic acid	$C_3H_6O_3$	90.0317	-0.47	P1_E
Uridine	$C_9H_{12}N_2O_6$	244.0695	-2.42	P1_E
Fructose bisphosphate	$C_6H_{14}O_{12}P_2$	339.9960	-3.01	P1_E
Carnosine	$C_9H_{14}N_4O_3$	226.1066	-4.48	P1_E
Nicotinamide	$C_6H_6N_2O$	122.0480	-0.39	P1_E
Shikimate	$C_7H_{10}O_5$	174.0528	-1.64	P1_E
Succinic acid	$C_4H_6O_4$	118.0266	-0.40	P1_F
Phenylalanine	$C_9H_{11}NO_2$	165.0790	-1.18	P1_F
Uracil	$C_4H_4N_2O_2$	112.0273	-0.86	P1_F
Malic acid	$C_4H_6O_5$	134.0215	-1.11	P1_F

Name	formula	exact mass	logP	Mix
Aspartic acid	$C_4H_7NO_4$	133.0375	-3.50	P1_F
dCMP	$C_9H_{14}N_3O_7P$	307.0569	-2.28	P1_F
Hypoxanthine	$C_5H_4N_4O$	136.0385	-0.05	P1_F
Creatine	$C_4H_9N_3O_2$	131.0695	-2.86	P1_F
Dopamine	$C_9H_{11}NO_4$	197.0688	-1.79	P1_F
Guanosine	$C_{10}H_{13}N_5O_5$	283.0917	-2.71	P1_F
Dihydrouracil	$C_4H_6N_2O_2$	114.0429	-1.21	P1_F
Malic acid	$C_4H_6O_5$	134.0215	-1.11	P1_F
Lysine	$C_6H_{14}N_2O_2$	146.1055	-3.21	P1_G
Tyrosine	$C_9H_{11}NO_3$	181.0739	-1.49	P1_G
Glycerol	$C_3H_8O_3$	92.0473	-1.84	P1_G
Asparagine	$C_4H_8N_2O_3$	132.0535	-4.29	P1_G
Valine	$C_5H_{11}NO_2$	117.0790	-1.95	P1_G
Guanine	$C_5H_5N_5O$	151.0494	-0.59	P1_G
Homoserine	$C_4H_9NO_3$	119.0582	-3.83	P1_G
Pyridoxine	$C_8H_{11}NO_3$	169.0739	-0.95	P1_G
dAMP	$C_{10}H_{14}N_5O_6P$	331.0682	-3.85	P1_G
Tartaric acid	$C_4H_6O_6$	150.0164	-1.83	P1_G
Nicotinamide mononucleotide	$C_{11}H_{15}N_2O_8P$	334.0566	-6.24	P1_G
Folic acid	$C_{19}H_{19}N_7O_6$	441.1397	-0.68	P1_G
Isocitric acid	$C_6H_8O_7$	192.0270	-1.45	P1_H
Thiourea	CH_4N_2S	76.0095	-0.47	P1_H
Diethanolamine	$C_4H_{11}NO_2$	105.0790	-1.57	P1_H

Name	formula	exact mass	logP	Mix
Aminoisobutanoic acid	$C_4H_9NO_2$	103.0633	-2.63	P1_H
Cys-Gly	$C_5H_{10}N_2O_3S$	178.0412	-3.90	P1_H
2-phosphoglyceric acid	$C_3H_7O_7P$	185.9929	-1.64	P1_H
Guanidinoacetic acid	$C_3H_7N_3O_2$	117.0538	-3.13	P1_H
Creatinine	$C_4H_7N_3O$	113.0589	-1.06	P1_H
N-Acetyltryptophan	$C_{13}H_{14}N_2O_3$	246.1004	1.00	P1_H
Trans-aconitic acid	$C_6H_6O_6$	174.0164	-0.52	P1_H
N-Acetylmannosamine	$C_8H_{15}NO_6$	221.0899	-3.22	P1_H
Glucose 6-phosphate	$C_6H_{13}O_9P$	260.0297	-3.06	P1_H
Diaminopimelic acid	$C_7H_{14}N_2O_4$	190.0954	-5.55	P2_A
Aminoadipic acid	$C_6H_{11}NO_4$	161.0688	-2.80	P2_A
Deoxycytidine	$C_9H_{13}N_3O_4$	227.0906	-1.90	P2_A
Noradrenaline	$C_8H_{11}NO_3$	169.0739	-0.68	P2_A
Glucosamine 6-phosphate	$C_6H_{14}NO_8P$	259.0457	-4.18	P2_A
Tartric acid	$C_4H_6O_6$	150.0164	-1.83	P2_A
3-Dehydroshikimic acid	$C_7H_8O_5$	172.0372	-1.02	P2_A
Norspermidine	$C_6H_{17}N_3$	131.1422	-1.67	P2_A
Homocysteine	$C_4H_9NO_2S$	135.0354	-2.58	P2_A
Theophylline	$C_7H_8N_4O_2$	180.0647	-0.77	P2_A
Leucine	$C_6H_{13}NO_2$	131.0946	-1.59	P2_A
Trehalose	$C_{12}H_{22}O_{11}$	342.1162	-4.70	P2_A
Betaine	$C_5H_{11}NO_2$	117.0790	-4.49	P2_B
Tryptophan	$C_{11}H_{12}N_2O_2$	204.0899	-1.09	P2_B

Name	formula	exact mass	logP	Mix
3-Sulfinoalanine	$C_3H_7NO_4S$	153.0096	-4.62	P2_B
O-Succinyl-homoserine	$C_8H_{13}NO_6$	219.0743	-3.61	P2_B
Allantoin	$C_4H_6N_4O_3$	158.0440	-2.36	P2_B
Glyceraldehyde	$C_3H_6O_3$	90.0317	-1.68	P2_B
D-Glucuronolactone	$C_6H_8O_6$	176.0321	-1.95	P2_B
(2-aminoethyl)phosphonate	$C_2H_8NO_3P$	125.0242	-3.17	P2_B
Selenomethionine	$C_5H_{11}NO_2Se$	196.9955	-3.30	P2_B
Maleimide	$C_4H_3NO_2$	97.0164	-0.64	P2_B
N,N-dimethyl-arginine	$C_8H_{18}N_4O_2$	202.1430	-2.67	P2_B
Glucosamine	$C_6H_{13}NO_5$	179.0794	-3.04	P2_B
Paraxanthine	$C_7H_8N_4O_2$	180.0647	0.24	P2_C
Adenosine 5'-diphosphate	$C_{10}H_{15}N_5O_{10}P_2$	427.0294	-4.71	P2_C
2-Deoxy-D-glucose	$C_6H_{12}O_5$	164.0685	-2.85	P2_C
1-methyl-l-histidine	$C_7H_{11}N_3O_2$	169.0851	-3.07	P2_C
Galactitol	$C_6H_{14}O_6$	182.0790	-3.73	P2_C
Oxoproline	$C_5H_7NO_3$	129.0426	-0.89	P2_C
4-pyridoxic acid	$C_8H_9NO_4$	183.0532	-0.75	P2_C
Quinolinic acid	$C_7H_5NO_4$	167.0219	-1.03	P2_C
Methylguanidine	$C_2H_7N_3$	73.0640	-0.96	P2_C
Caffeine	$C_8H_{10}N_4O_2$	194.0804	-0.55	P2_C
3-hydroxy-3-methylglutaryl-CoA	$C_{27}H_{44}N_7O_{20}P_3S$	911.1575	-6.41	P2_C
Glucuronic acid	$C_6H_{10}O_7$	194.0427	-2.61	P2_C
1-methyladenosine	$C_{11}H_{15}N_5O_4$	281.1124	-2.09	P2_D

Name	formula	exact mass	logP	Mix
Deoxyuridine	$C_9H_{12}N_2O_5$	228.0746	-1.51	P2_D
Trans-4-hydroxy-l-proline	$C_5H_9NO_3$	131.0582	-3.72	P2_D
Urocanic acid	$C_6H_6N_2O_2$	138.0429	-1.01	P2_D
Kynurenine	$C_{10}H_{12}N_2O_3$	208.0848	-1.91	P2_D
Pyroglutamic acid	$C_5H_7NO_3$	129.0426	-0.89	P2_D
4-acetamidobutanoic acid	$C_6H_{11}NO_3$	145.0739	-0.80	P2_D
Trans-cyclohexanediol	$C_6H_{12}O_2$	116.0837	0.21	P2_D
Melanin	$C_{18}H_{10}N_2O_4$	318.0641	-0.27	P2_D
Dopamine	$C_8H_{11}NO_2$	153.0790	0.03	P2_D
Putrescine	$C_4H_{12}N_2$	88.1000	-0.85	P2_D
Lysine	$C_6H_{14}N_2O_2$	146.1055	-3.21	P2_D
Citicoline	$C_{14}H_{26}N_4O_{11}P_2$	488.1073	-7.11	P2_E
1,3-diaminopropane	$C_3H_{10}N_2$	74.0844	-1.36	P2_E
Phosphoserine	$C_3H_8NO_6P$	185.0089	-3.18	P2_E
1-aminocyclopropanecarboxylic acid	$C_4H_7NO_2$	101.0477	-2.72	P2_E
Glutarylcarnitine	$C_{12}H_{21}NO_6$	275.1369	-4.18	P2_E
Cystathionine	$C_7H_{14}N_2O_4S$	222.0674	-5.82	P2_E
Norvaline	$C_5H_{11}NO_2$	117.0790	-1.87	P2_E
3-hydroxymethylglutaric acid	$C_6H_{10}O_5$	162.0528	-0.75	P2_E
Phosphonoacetic acid	$C_2H_5O_5P$	139.9875	-1.60	P2_E
Picolinic acid	$C_6H_5NO_2$	123.0320	-0.65	P2_E
Ethanolamine	C_2H_7NO	61.0528	-1.32	P2_E
Arginine	$C_6H_{14}N_4O_2$	174.1117	-3.24	P2_E

Name	formula	exact mass	logP	Mix
trans-4-hydroxy-L-proline	$C_5H_9NO_3$	131.0582	-3.72	P2_F
Fucose	$C_6H_{12}O_5$	164.0685	-1.89	P2_F
Homocystine	$C_8H_{16}N_2O_4S_2$	268.0551	-5.32	P2_F
N-Methylglutamic acid	$C_6H_{11}NO_4$	161.0688	-3.00	P2_F
D-Ornithine	$C_5H_{12}N_2O_2$	132.0899	-3.66	P2_F
Xanthosine	$C_{10}H_{12}N_4O_6$	284.0757	-1.19	P2_F
3-methylcrotonyl-CoA	$C_{26}H_{42}N_7O_{17}P_3S$	849.1571	-4.26	P2_F
Thyrotropin releasing hormone	$C_{16}H_{22}N_6O_4$	362.1703	-3.27	P2_F
Cytidine monophosphate	$C_9H_{14}N_3O_8P$	323.0519	-3.15	P2_F
N-Methylaspartic acid	$C_5H_9NO_4$	147.0532	-3.27	P2_F
Galactaric acid	$C_6H_{10}O_8$	210.0376	-3.09	P2_F
Histidine	$C_6H_9N_3O_2$	155.0695	-3.29	P2_F
Nicotinic acid adenine dinucleotide phosphate	$C_{21}H_{27}N_6O_{18}P_3$	744.0595	-9.60	P2_G
N-Acetylasparagine	$C_6H_{10}N_2O_4$	174.0641	-2.21	P2_G
Pipecolic acid	$C_6H_{11}NO_2$	129.0790	-2.12	P2_G
Glucose 6-phosphate	$C_6H_{13}O_9P$	260.0297	-3.06	P2_G
NADP	$C_{21}H_{28}N_7O_{17}P_3$	743.0755	-10.40	P2_G
Carbamoyl phosphate	CH_4NO_5P	140.9827	-1.21	P2_G
Isopentenyl pyrophosphate	$C_5H_{12}O_7P_2$	246.0058	0.20	P2_G
GTP	$C_{10}H_{16}N_5O_{14}P_3$	522.9907	-3.67	P2_G
dDTP-D-glucose	$C_{16}H_{26}N_2O_{16}P_2$	564.0758	-3.70	P2_G
Agmatine sulfate	$C_5H_{14}N_4$	130.1218	-1.23	P2_G
Glycolaldehyde	$C_2H_4O_2$	60.0211	-1.20	P2_G

Name	formula	exact mass	logP	Mix
dGTP	$C_{10}H_{16}N_5O_{13}P_3$	506.9957	-3.23	P2_G
N-acetylglycine	$C_4H_7NO_3$	117.0426	-1.33	P2_H
N-acetylaspartic acid	$C_6H_9NO_5$	175.0481	-1.40	P2_H
IDP	$C_{10}H_{14}N_4O_{11}P_2$	428.0134	-3.58	P2_H
Palmitoylcarnitine	$C_{23}H_{45}NO_4$	399.3349	2.03	P2_H
dGMP	$C_{10}H_{14}N_5O_7P$	347.0631	-2.23	P2_H
Nicotinamide hypoxanthine dinucleotide	$C_{21}H_{26}N_6O_{15}P_2$	664.0931	-7.73	P2_H
S-Adenosylmethionine	$C_{15}H_{22}N_6O_5S$	398.1372	-5.32	P2_H
6-phosphogluconic acid	$C_6H_{13}O_{10}P$	276.0246	-3.53	P2_H
α-hydroxyisobutyric acid	$C_4H_8O_3$	104.0473	-0.04	P2_H
Cysteic acid	$C_3H_7NO_5S$	169.0045	-2.96	P2_H
Adenosine-monophosphate	$C_{10}H_{14}N_5O_7P$	347.0631	-4.74	P2_H
Gluconic acid	$C_6H_{12}O_7$	196.0583	-3.41	P2_H
Putrescine	$C_4H_{12}N_2$	88.1000	-0.85	P3_A
Deoxycarnitine	$C_7H_{15}NO_2$	145.1103	-3.97	P3_A
Adenosine 2',3'-cyclic phosphate	$C_{10}H_{12}N_5O_6P$	329.0525	-3.45	P3_A
Mevalolactone	$C_6H_{10}O_3$	130.0630	-0.43	P3_A
Uridine diphosphate glucose	$C_{15}H_{24}N_2O_{17}P_2$	566.0550	-5.00	P3_A
Gamma,gamma-dimethylallyl pyrophosphate	$C_5H_{12}O_7P_2$	246.0058	0.30	P3_A
Deoxyuridine triphosphate	$C_9H_{15}N_2O_{14}P_3$	467.9736	-2.48	P3_A
Phosphorylcholine	$C_5H_{14}NO_4P$	183.0660	-4.79	P3_A
Uridine triphosphate	$C_9H_{15}N_2O_{15}P_3$	483.9685	-3.38	P3_A
6-hydroxydopamine	$C_8H_{11}NO_3$	169.0739	-0.15	P3_A

Name	formula	exact mass	logP	Mix
Thiamine	$C_{12}H_{17}N_4OS$	265.1118	-3.10	P3_A
dGDP	$C_{10}H_{15}N_5O_{10}P_2$	427.0294	-2.52	P3_A
5-methylcytosine	$C_5H_7N_3O$	125.0589	-0.75	P3_B
Glyceric acid	$C_3H_6O_4$	106.0266	-1.52	P3_B
Cytidine 2',3'-cyclic phosphate	$C_9H_{12}N_3O_7P$	305.0413	-2.12	P3_B
N,N,N-trimethyllysine	$C_9H_{20}N_2O_2$	188.1525	-6.20	P3_B
Phosphoenolpyruvic acid	$C_3H_5O_6P$	167.9824	-0.64	P3_B
Uridine diphosphate glucose	$C_{15}H_{24}N_2O_{17}P_2$	566.0550	-5.00	P3_B
Galactose 1-phosphate	$C_6H_{13}O_9P$	260.0297	-3.06	P3_B
Pyridoxal-phosphate	$C_8H_{10}NO_6P$	247.0246	-2.09	P3_B
Dihydroxyacetone phosphate	$C_3H_7O_6P$	169.9980	-1.65	P3_B
Phosphoenolpyruvic acid	$C_3H_5O_6P$	167.9824	-0.64	P3_B
Mannose 6-phosphate	$C_6H_{13}O_9P$	260.0297	-3.06	P3_B
3-phosphoglyceric acid	$C_3H_7O_7P$	185.9929	-1.64	P3_B
L-carnitine	$C_7H_{15}NO_3$	161.1052	-4.89	P3_C
O-phosphoethanolamine	$C_2H_8NO_4P$	141.0191	-2.45	P3_C
O-Acetylserine	$C_5H_9NO_4$	147.0532	-3.45	P3_C
Thymidine-monophosphate	$C_{10}H_{15}N_2O_8P$	322.0566	-1.24	P3_C
Cyclic AMP	$C_{10}H_{12}N_5O_6P$	329.0525	-3.39	P3_C
ADP-glucose	$C_{16}H_{25}N_5O_{15}P_2$	589.0822	-6.77	P3_C
Fructose 6-phosphate	$C_6H_{13}O_9P$	260.0297	-3.39	P3_C
Adenosine 3',5'-diphosphate	$C_{10}H_{15}N_5O_{10}P_2$	427.0294	-4.80	P3_C
3-Nitro-L-tyrosine	$C_9H_{10}N_2O_5$	226.0590	-1.55	P3_C

Name	formula	exact mass	logP	Mix
P-Octopamine	$C_8H_{11}NO_2$	153.0790	-0.32	P3_C
N-α-Acetyllysine	$C_8H_{16}N_2O_3$	188.1161	-3.15	P3_C
Uridine diphosphategalactose	$C_{15}H_{24}N_2O_{17}P_2$	566.0550	-5.00	P3_C
Spermidine	$C_7H_{19}N_3$	145.1579	-1.15	P3_D
Pyridoxamine	$C_8H_{12}N_2O_2$	168.0899	-1.61	P3_D
5-aminolevulinic acid	$C_5H_9NO_3$	131.0582	-3.25	P3_D
Deoxyuridine-monophosphate	$C_9H_{13}N_2O_8P$	308.0410	-1.64	P3_D
ATP	$C_{10}H_{16}N_5O_{13}P_3$	506.9957	-5.80	P3_D
Ribose 1,5-bisphosphate	$C_5H_{12}O_{11}P_2$	309.9855	-2.55	P3_D
Xanthosine-monophosphate	$C_{10}H_{13}N_4O_9P$	364.0420	-2.18	P3_D
FAD	$C_{27}H_{33}N_9O_{15}P_2$	785.1571	-4.68	P3_D
Deoxyguanosine	$C_{10}H_{13}N_5O_4$	267.0968	-1.81	P3_D
Orotic acid	$C_5H_4N_2O_4$	156.0171	-1.23	P3_D
Lauroylcarnitine	$C_{19}H_{37}NO_4$	343.2723	0.26	P3_D
1-methylnicotinamide	$C_7H_9N_2O$	137.0709	-4.34	P3_D
Spermine	$C_{10}H_{26}N_4$	202.2157	-1.45	P3_E
N-Acetylmethionine	$C_7H_{13}NO_3S$	191.0616	-0.11	P3_E
Carbamoyl phosphate	CH_4NO_5P	140.9827	-1.21	P3_E
Phosphoribosyl pyrophosphate	$C_5H_{13}O_{14}P_3$	389.9518	-2.97	P3_E
AICAR	$C_9H_{15}N_4O_8P$	338.0628	-4.81	P3_E
Uridine diphosphate-N-acetylgalactosamine	$C_{17}H_{27}N_3O_{17}P_2$	607.0816	-5.28	P3_E
Glyceraldehyde 3-phosphate	$C_3H_7O_6P$	169.9980	-1.80	P3_E
cGMP	$C_{10}H_{12}N_5O_7P$	345.0474	-2.09	P3_E

Name	formula	exact mass	logP	Mix
Homocysteine thiolactone	C_4H_7NOS	117.0248	-0.25	P3_E
O-phosphoserine	$C_3H_8NO_6P$	185.0089	-3.18	P3_E
S-adenosylhomocysteine	$C_{14}H_{20}N_6O_5S$	384.1216	-4.02	P3_E
L-Ornithine	$C_5H_{12}N_2O_2$	132.0899	-3.66	P3_E
Adenine	$C_5H_5N_5$	135.0545	-0.57	P3_F
Normetanephrine	$C_9H_{13}NO_3$	183.0895	-0.39	P3_F
Uridine diphosphate-n-acetylglucosamine	$C_{17}H_{27}N_3O_{17}P_2$	607.0816	-5.28	P3_F
Guanosine diphosphate	$C_{10}H_{15}N_5O_{11}P_2$	443.0243	-3.42	P3_F
Phosphocreatine	$C_4H_{10}N_3O_5P$	211.0358	-2.25	P3_F
Uridine diphosphate glucuronic acid	$C_{15}H_{22}N_2O_{18}P_2$	580.0343	-4.68	P3_F
N,N-dimethylarginine	$C_8H_{18}N_4O_2$	202.1430	-2.67	P3_F
Cytidine diphosphate	$C_9H_{15}N_3O_{11}P_2$	403.0182	-3.34	P3_F
Selenocystamine	$C_4H_{12}N_2Se_2$	247.9331	-1.70	P3_F
Histamine	$C_5H_9N_3$	111.0796	-0.70	P3_F
Indoxyl sulfate	$C_8H_7NO_4S$	213.0096	1.29	P3_F
Ethyl 3-ureidopropionic acid	$C_6H_{12}N_2O_3$	160.0848	-0.92	P3_F
Deoxyribose	$C_5H_{10}O_4$	134.0579	-1.40	P3_G
Phytic acid	$C_6H_{18}O_{24}P_6$	659.8614	-4.52	P3_G
Thiamine monophosphate	$C_{12}H_{17}N_4O_4PS$	344.0708	-5.70	P3_G
Uracil 5-carboxylic acid	$C_5H_4N_2O_4$	156.0171	-1.27	P3_G
S-hexyl-glutathione	$C_{16}H_{29}N_3O_6S$	391.1777	-2.38	P3_G
Glyoxylic acid	$C_2H_2O_3$	74.0004	-0.13	P3_G
GMP	$C_{10}H_{14}N_5O_8P$	363.0580	-3.13	P3_G

Name	formula	exact mass	logP	Mix
N-Acetylalanine	$C_5H_9NO_3$	131.0582	-0.76	P3_G
4-guanidinobutanoic acid	$C_5H_{11}N_3O_2$	145.0851	-2.65	P3_G
Hydroxypyruvic acid	$C_3H_4O_4$	104.0110	-0.75	P3_G
D-mannosamine	$C_6H_{13}NO_5$	179.0794	-3.04	P3_G
Cytochrome C	$C_{42}H_{52}FeN_8O_6S_2$	884.2801	0.76	P3_G
O-Acetylcarnitine	$C_9H_{17}NO_4$	203.1158	-4.45	P3_H
Riboflavin	$C_{17}H_{20}N_4O_6$	376.1383	-0.92	P3_H
Methyl galactoside	$C_7H_{14}O_6$	194.0790	-2.29	P3_H
Glutaric acid	$C_5H_8O_4$	132.0423	0.05	P3_H
Dihydroxyfumaric acid	$C_4H_4O_6$	148.0008	-0.01	P3_H
Cytidine monophosphate	$C_9H_{14}N_3O_8P$	323.0519	-3.15	P3_H
Guanosine diphosphate mannose	$C_{16}H_{25}N_5O_{16}P_2$	605.0772	-5.29	P3_H
5'-deoxyadenosine	$C_{10}H_{13}N_5O_3$	251.1018	-1.04	P3_H
Glutathione reduced	$C_{10}H_{17}N_3O_6S$	307.0838	-4.88	P3_H
Erythritol	$C_4H_{10}O_4$	122.0579	-2.47	P3_H
Glucosaminic acid	$C_6H_{13}NO_6$	195.0743	-5.78	P3_H
UTP	$C_9H_{15}N_2O_{15}P_3$	483.9685	-3.38	P3_H
Deoxyadenosine	$C_{10}H_{13}N_5O_3$	251.1018	-1.19	P4_A
N-Acetylputrescine	$C_6H_{14}N_2O$	130.1106	-1.03	P4_A
N-Acetylgalactosamine	$C_8H_{15}NO_6$	221.0899	-3.22	P4_A
N-Acetylglutamic acid	$C_7H_{11}NO_5$	189.0637	-1.11	P4_A
2,4-dihydroxypteridine	$C_6H_4N_4O_2$	164.0334	-0.06	P4_A
6-hydroxynicotinic acid	$C_6H_5NO_3$	139.0269	0.70	P4_A

Name	formula	exact mass	logP	Mix
N-Acetylcysteine	$C_5H_9NO_3S$	163.0303	-0.71	P4_A
IMP	$C_{10}H_{13}N_4O_8P$	348.0471	-2.93	P4_A
Pantothenic acid	$C_9H_{17}NO_5$	219.1107	-1.36	P4_A
2-aminoisobutyric acid	$C_4H_9NO_2$	103.0633	-2.41	P4_A
Aniline-2-sulfonate	$C_6H_7NO_3S$	173.0147	0.10	P4_A
S-carboxymethylcysteine	$C_5H_9NO_4S$	179.0252	-3.30	P4_A
Rhamnose	$C_6H_{12}O_5$	164.0685	-1.89	P4_B
Thiamine pyrophosphate	$C_{12}H_{18}N_4O_7P_2S$	424.0371	-5.80	P4_B
Histidinol	$C_6H_{11}N_3O$	141.0902	-1.67	P4_B
Thymidine-monophosphate	$C_{10}H_{15}N_2O_8P$	322.0566	-1.24	P4_B
Ureidopropionic acid	$C_4H_8N_2O_3$	132.0535	-1.43	P4_B
5-aminopentanoic acid	$C_5H_{11}NO_2$	117.0790	-2.44	P4_B
Norleucine	$C_6H_{13}NO_2$	131.0946	-1.43	P4_B
N-formylglycine	$C_3H_5NO_3$	103.0269	-1.38	P4_B
Adenosine	$C_{10}H_{13}N_5O_4$	267.0968	-2.09	P4_B
Raffinose	$C_{18}H_{32}O_{16}$	504.1690	-6.30	P4_B
Meso-tartric acid	$C_4H_6O_6$	168.0270	-1.83	P4_B
2-acetamido-2-deoxy-β-D-glucosylamine	$C_8H_{16}N_2O_5$	220.1059	-3.33	P4_B
Saccharic acid	$C_6H_{10}O_8$	210.0376	-3.09	P4_C
ATP	$C_{10}H_{16}N_5O_{13}P_3$	506.9957	-5.80	P4_C
3-methoxytyrosine	$C_{10}H_{13}NO_4$	211.0845	-1.65	P4_C
Lactose	$C_{12}H_{22}O_{11}$	342.1162	-4.70	P4_C
3-hydroxybutanoic acid	$C_4H_8O_3$	104.0473	-0.39	P4_C

Name	formula	exact mass	logP	Mix
4-imidazoleacetic acid	$C_5H_6N_2O_2$	126.0429	-1.40	P4_C
Galacturonic acid	$C_6H_{10}O_7$	194.0427	-2.61	P4_C
CTP	$C_9H_{16}N_3O_{14}P_3$	482.9845	-4.10	P4_C
cAMP	$C_{10}H_{12}N_5O_6P$	329.0525	-3.39	P4_C
Methionine sulfoximine	$C_5H_{12}N_2O_3S$	180.0569	-4.45	P4_C
Cis-4-hydroxy-D-proline	$C_5H_9NO_3$	131.0582	-3.72	P4_C
N1-Acetylspermine	$C_{12}H_{28}N_4O$	244.2263	-1.64	P4_C
Mesoxalic caid	$C_3H_2O_5$	117.9902	0.03	P4_D
NADPH	$C_{21}H_{30}N_7O_{17}P_3$	745.0911	-6.38	P4_D
3-methylhistamine	$C_6H_{11}N_3$	125.0953	-0.82	P4_D
Maleamic acid	$C_4H_5NO_3$	115.0269	-0.85	P4_D
Choline	$C_5H_{14}NO$	104.1070	-4.66	P4_D
Methyl 4-aminobutyric acid	$C_5H_{11}NO_2$	117.0790	-0.48	P4_D
N-formyl-L-methionine	$C_6H_{11}NO_3S$	177.0460	-0.16	P4_D
Acetylcholine	$C_7H_{16}NO_2$	146.1176	-4.22	P4_D
Oxalic acid	$C_2H_2O_4$	89.9953	-0.26	P4_D
5-hydroxytryptophan	$C_{11}H_{12}N_2O_3$	220.0848	-1.39	P4_D
D-alanine	$C_3H_7NO_2$	89.0477	-2.84	P4_D
Theobromine	$C_7H_8N_4O_2$	180.0647	-0.77	P4_D
Guanidinosuccinic acid	$C_5H_9N_3O_4$	175.0593	-3.27	P4_E
Histidine	$C_6H_9N_3O_2$	155.0695	-3.29	P4_E
Allothreonine	$C_4H_9NO_3$	119.0582	-3.47	P4_E
Phosphocreatine	$C_4H_{10}N_3O_5P$	211.0358	-2.25	P4_E

Name	formula	exact mass	logP	Mix
Spermidine	$C_7H_{19}N_3$	145.1579	-1.15	P4_E
Adenosine diphosphate ribose	$C_{15}H_{23}N_5O_{14}P_2$	559.0717	-6.07	P4_E
2-methoxyethanol	$C_3H_8O_2$	76.0524	-0.57	P4_E
Citramalic acid	$C_5H_8O_5$	148.0372	-0.68	P4_E
Anserine	$C_{10}H_{16}N_4O_3$	240.1222	-4.27	P4_E
Biliverdin	$C_{33}H_{34}N_4O_6$	582.2478	3.17	P4_E
5-hydroxylysine	$C_6H_{14}N_2O_3$	162.1004	-4.41	P4_E
Cysteamine	C_2H_7NS	77.0299	-0.42	P4_E
Ophthalmic acid	$C_{11}H_{19}N_3O_6$	289.1274	-4.40	P4_F
2,3-diaminopropionic acid	$C_3H_8N_2O_2$	104.0586	-4.02	P4_F
Trigonelline	$C_7H_7NO_2$	137.0477	-3.53	P4_F
Epinephrine	$C_9H_{13}NO_3$	183.0895	-0.43	P4_F
3,4-dihydroxyphenylglycol	$C_8H_{10}O_4$	170.0579	-0.03	P4_F
Cadaverine	$C_5H_{14}N_2$	102.1157	-0.40	P4_F
2-hydroxybutyric acid	$C_4H_8O_3$	104.0473	0.05	P4_F
Coenzyme A	$C_{21}H_{36}N_7O_{16}P_3S$	767.1152	-5.72	P4_F
Oxalomalic acid	$C_6H_6O_8$	206.0063	-1.09	P4_F
ITP	$C_{10}H_{15}N_4O_{14}P_3$	507.9798	-2.69	P4_F
CDP-ethanolamine	$C_{11}H_{20}N_4O_{11}P_2$	446.0604	-4.98	P4_F
2,5-dimethylpyrazine	$C_6H_8N_2$	108.0687	-0.20	P4_F
Stachyose	$C_{24}H_{42}O_{21}$	666.2219	-8.07	P4_G
Deoxycytidine-diphosphate	$C_9H_{15}N_3O_{10}P_2$	387.0233	-2.59	P4_G
2,3-butanediol	$C_4H_{10}O_2$	90.0681	-0.38	P4_G

Name	formula	exact mass	logP	Mix
D-ribose 5-phosphate	$C_5H_{11}O_8P$	230.0192	-2.43	P4_G
Hydroxykynurenine	$C_{10}H_{12}N_2O_4$	224.0797	-2.21	P4_G
Galactosamine	$C_6H_{13}NO_5$	179.0794	-3.04	P4_G
Deoxyadenosine triphosphate	$C_{10}H_{16}N_5O_{12}P_3$	491.0008	-4.91	P4_G
Glycerol 3-phosphate	$C_3H_9O_6P$	172.0137	-1.96	P4_G
Cyanocobalamin	$C_{63}H_{89}CoN_{14}O_{14}P$	1354.5674	-2.18	P4_G
4-hydroxy-L-phenylglycine	$C_8H_9NO_3$	167.0582	-1.78	P4_G
N-Acetylserine	$C_5H_9NO_4$	147.0532	-1.81	P4_G
Uridine 5'-diphosphate	$C_9H_{14}N_2O_{12}P_2$	404.0022	-2.96	P4_G
Beta-glycerophosphate	$C_3H_9O_6P$	172.0137	-1.96	P4_H
Glucose 1-phosphate	$C_6H_{13}O_9P$	260.0297	-3.06	P4_H
Glucosamine 6-sulfate	$C_6H_{13}NO_8S$	259.0362	-3.99	P4_H
Methyglutaric acid	$C_6H_{10}O_4$	146.0579	0.33	P5_A
Sorbic acid	$C_6H_8O_2$	112.0524	1.45	P5_A
Monoethylmalonic acid	$C_5H_8O_4$	132.0423	0.17	P5_A
Gluconolactone	$C_6H_{10}O_6$	178.0472	1.22	P5_A
4-hydroxybenzoic acid	$C_7H_6O_3$	138.0317	1.33	P5_A
Tyramine	$C_8H_{11}NO$	137.0841	0.68	P5_A
Cortisol	$C_{21}H_{30}O_5$	362.2093	1.28	P5_A
Prenol	$C_5H_{10}O$	86.0732	0.84	P5_A
3-hydroxybenzaldehyde	$C_7H_6O_2$	122.0368	1.38	P5_A
Xanthurenic acid	$C_{10}H_7NO_4$	205.0375	-0.17	P5_A
2-methylpropanal	C_4H_8O	72.0570	0.87	P5_A

Name	formula	exact mass	logP	Mix
Propionic acid	$C_3H_6O_2$	74.0368	0.48	P5_A
Trimethylamine	C_3H_9N	59.0735	0.19	P5_B
Melatonin	$C_{13}H_{16}N_2O_2$	232.1212	1.15	P5_B
Maleic acid	$C_4H_4O_4$	116.0110	-0.04	P5_B
Pentanoic acid	$C_5H_{10}O_2$	102.0681	1.37	P5_B
Propanoic acid	$C_3H_6O_2$	74.0368	0.48	P5_B
Bilirubin	$C_{33}H_{36}N_4O_6$	584.2635	3.12	P5_B
Nicotine	$C_{10}H_{14}N_2$	162.1157	1.16	P5_B
Pregnenolone sulfate	$C_{21}H_{32}O_5S$	396.1970	3.64	P5_B
Kynurenic acid	$C_{10}H_7NO_3$	189.0426	1.58	P5_B
Isobutyric acid	$C_4H_8O_2$	88.0524	1.02	P5_B
3-hydroxybenzyl alcohol	$C_7H_8O_2$	124.0524	0.90	P5_B
Aniline	C_6H_7N	93.0578	1.14	P5_B
Acetoin	$C_4H_8O_2$	88.0519	1.31	P5_C
3,5-diiodo-L-tyrosine	$C_9H_9I_2NO_3$	432.8672	0.37	P5_C
Mandelic acid	$C_8H_8O_3$	152.0473	0.90	P5_C
Tryptamine	$C_{10}H_{12}N_2$	160.1000	1.49	P5_C
Benzoic acid	$C_7H_6O_2$	122.0368	1.63	P5_C
Glutaric acid	$C_5H_8O_4$	132.0423	0.05	P5_C
Indole-3-acetic acid	$C_{10}H_9NO_2$	175.0633	1.71	P5_C
Caffeic acid	$C_9H_8O_4$	180.0423	1.53	P5_C
Lumichrome	$C_{12}H_{10}N_4O_2$	242.0804	2.73	P5_C
β-alanine	$C_3H_7NO_2$	89.0477	-3.17	P5_C

Name	formula	exact mass	logP	Mix
N-Acetylphenylalanine	$C_{11}H_{13}NO_3$	207.0895	0.90	P5_C
N-Acetylproline	$C_7H_{11}NO_3$	157.0739	-0.49	P5_C
L-Tryptophanamide	$C_{11}H_{13}N_3O$	203.1059	0.37	P5_D
Phenol	C_6H_6O	94.0419	1.67	P5_D
N-Methyltryptamine	$C_{11}H_{14}N_2$	174.1157	1.92	P5_D
Oxaloacetic acid	$C_4H_4O_5$	132.0059	-0.04	P5_D
2,3-dihydroxybenzoica cid	$C_7H_6O_4$	154.0266	1.67	P5_D
2-propenoic acid	$C_3H_4O_2$	72.0211	0.53	P5_D
Indole-3-ethanol	$C_{10}H_{11}NO$	161.0841	1.59	P5_D
Ferulic acid	$C_{10}H_{10}O_4$	194.0579	1.67	P5_D
Glycocholic acid	$C_{26}H_{43}NO_6$	465.3090	1.38	P5_D
Phenylethanolamine	$C_8H_{11}NO$	137.0841	0.47	P5_D
Thiopurine S-methylether	$C_6H_6N_4S$	166.0313	0.88	P5_D
2-hydroxy-4-(methylthio)butanoic acid	$C_5H_{10}O_3S$	150.0351	0.18	P5_D
Glycochenodeoxycholic acid	$C_{26}H_{43}NO_5$	449.3141	2.61	P5_E
Benzoic acid	$C_7H_6O_2$	122.0368	1.63	P5_E
3-amino-5-hydroxybenzoic acid	$C_7H_7NO_3$	153.0426	0.34	P5_E
Pyrocatechol	$C_6H_6O_2$	110.0368	1.37	P5_E
3,4-dihydroxybenzoic acid	$C_7H_6O_4$	154.0266	1.02	P5_E
Cyclopentanone	C_5H_8O	84.0575	1.04	P5_E
Pantolactone	$C_6H_{10}O_3$	130.0630	0.18	P5_E
Guaiacol	$C_7H_8O_2$	124.0524	1.51	P5_E
2-hydroxyphenylacetic acid	$C_8H_8O_3$	152.0473	1.31	P5_E

Name	formula	exact mass	logP	Mix
10-hydroxydecanoic acid	$C_{10}H_{20}O_3$	188.1412	2.15	P5_E
Didecanoyl-glycerophosphocholine	$C_{28}H_{56}NO_8P$	565.3744	2.78	P5_E
2-hydroxypyridine	C_5H_5NO	95.0371	1.05	P5_E
3,4-dihydroxyphenylacetic acid	$C_8H_8O_4$	168.0423	1.00	P5_F
N6-(delta2-isopentenyl)-adenine	$C_{10}H_{13}N_5$	203.1171	1.09	P5_F
Methyl vanillic acid	$C_9H_{10}O_4$	182.0579	1.52	P5_F
2-oxobutanoic acid	$C_4H_6O_3$	102.0317	0.77	P5_F
Lipoamide	$C_8H_{15}NOS_2$	205.0595	1.31	P5_F
3-hydroxyanthranilic acid	$C_7H_7NO_3$	153.0426	1.15	P5_F
3-(4-hydroxyphenyl)pyruvic acid	$C_9H_8O_4$	180.0423	1.60	P5_F
Hexanoic acid	$C_6H_{12}O_2$	116.0837	1.81	P5_F
Methylmalonic acid	$C_4H_6O_4$	118.0266	0.21	P5_F
Serotonin	$C_{10}H_{12}N_2O$	176.0950	0.48	P5_F
Cortisol 21-acetate	$C_{23}H_{32}O_6$	404.2199	1.72	P5_F
Indole-3-acetamide	$C_{10}H_{10}N_2O$	174.0793	-0.92	P5_F
Hippuric acid	$C_9H_9NO_3$	179.0582	0.53	P5_G
Ethylmalonic acid	$C_5H_8O_4$	132.0423	0.66	P5_G
3,5-diiodo-L-thyronine	$C_{15}H_{13}I_2NO_4$	524.8934	1.87	P5_G
Fumaric acid	$C_4H_4O_4$	116.0110	-0.04	P5_G
Benzaldehyde	C_7H_6O	106.0419	1.69	P5_G
4-hydroxybenzaldehyde	$C_7H_6O_2$	122.0368	1.38	P5_G
3-(2-hydroxyphenyl)propanoic acid	$C_9H_{10}O_3$	166.0630	1.75	P5_G
3-methoxytyramine	$C_9H_{13}NO_2$	167.0946	0.53	P5_G

Name	formula	exact mass	logP	Mix
Benzylamine	C_7H_9N	107.0735	1.10	P5_G
2-quinolinecarboxylic acid	$C_{10}H_7NO_2$	173.0477	0.79	P5_G
Serotonin	$C_{10}H_{12}N_2O$	176.0950	0.48	P5_G
Pterin	$C_6H_5N_5O$	163.0494	-0.96	P5_G
4-aminobenzoic acid	$C_7H_7NO_2$	137.0477	0.80	P5_H
2-aminophenol	C_6H_7NO	109.0528	0.84	P5_H
6-carboxyhexanoic acid	$C_7H_{12}O_4$	160.0736	0.94	P5_H
Indole-3-pyruvic acid	$C_{11}H_9NO_3$	203.0582	2.00	P5_H
Dehydroascorbic acid	$C_6H_6O_6$	174.0164	-0.67	P5_H
3-amino-4-hydroxybenzoic acid	$C_7H_7NO_3$	153.0426	0.35	P5_H
3,4 dihydroxymandelic acid	$C_8H_8O_5$	184.0372	0.29	P5_H
Anthranilic acid	$C_7H_7NO_2$	137.0477	1.45	P5_H
Dihydrobiopterin	$C_9H_{13}N_5O_3$	239.1018	-2.29	P5_H
Butanoic acid	$C_4H_8O_2$	88.0524	0.92	P5_H
Indole-3-acetic acid	$C_{10}H_9NO_2$	175.0633	1.71	P5_H
5-valerolactone	$C_5H_8O_2$	100.0524	0.59	P5_H
2,5-dihydroxybenzoic acid	$C_7H_6O_4$	154.0266	1.67	P6_A
4-quinolinecarboxylic acid	$C_{10}H_7NO_2$	173.0477	1.63	P6_A
Hydroquinone	$C_6H_6O_2$	110.0368	1.37	P6_A
Dethiobiotin	$C_{10}H_{18}N_2O_3$	214.1317	0.73	P6_A
3-methyl-2-oxovaleric acid	$C_6H_{10}O_3$	130.0630	1.75	P6_A
Oxoglutaric acid	$C_5H_6O_5$	146.0215	-0.11	P6_A
N-Acetylserotonin	$C_{12}H_{14}N_2O_2$	218.1055	1.00	P6_A

Name	formula	exact mass	logP	Mix
Allyl isothiocyanate	C_4H_5NS	99.0137	1.05	P6_A
Itaconic acid	$C_5H_6O_4$	130.0266	0.05	P6_A
Azelaic acid	$C_9H_{16}O_4$	188.1049	1.82	P6_A
Oxoadipic acid	$C_6H_8O_5$	160.0372	0.34	P6_A
2-methylglutaric acid	$C_6H_{10}O_4$	146.0579	0.59	P6_A
Phenylacetaldehyde	C_8H_8O	120.0575	1.45	P6_B
2-methylbutanal	$C_5H_{10}O$	86.0732	1.31	P6_B
Phenyl acetic acid	$C_8H_8O_2$	136.0524	1.58	P6_B
Diacetyl	$C_4H_6O_2$	86.0368	0.40	P6_B
Pyruvic acid	$C_3H_4O_3$	88.0160	0.07	P6_B
Trans-cinnamaldehyde	C_9H_8O	132.0575	1.98	P6_B
2,6-dihydroxypyridine	$C_5H_5NO_2$	111.0320	0.15	P6_B
Phenethylamine	$C_8H_{11}N$	121.0891	1.39	P6_B
Methyl acetoacetic acid	$C_5H_8O_3$	116.0473	0.14	P6_B
Suberic acid	$C_8H_{14}O_4$	174.0892	1.38	P6_B
Adipic acid	$C_6H_{10}O_4$	146.0579	0.49	P6_B
Geranyl-PP	$C_{10}H_{20}O_7P_2$	314.0684	1.96	P6_B
N-Acetylleucine	$C_8H_{15}NO_3$	173.1052	0.49	P6_C
2',4'-dihydroxyacetophenone	$C_8H_8O_3$	152.0473	1.57	P6_C
Benzyl alcohol	C_7H_8O	108.0575	1.21	P6_C
Monomethylglutaric acid	$C_6H_{10}O_4$	146.0579	0.19	P6_C
Indole-3-methyl acetic acid	$C_{11}H_{11}NO_2$	189.0790	1.86	P6_C
Mevalonic acid	$C_6H_{12}O_4$	148.0730	0.18	P6_C

Name	formula	exact mass	logP	Mix
3-methoxy-4-hydroxymandelic acid	$C_9H_{10}O_5$	198.0528	0.43	P6_C
Homovanillic acid	$C_9H_{10}O_4$	182.0579	1.15	P6_C
2-methylmaleic acid	$C_5H_6O_4$	130.0266	0.35	P6_C
1-phenylethanol	$C_8H_{10}O$	122.0732	1.62	P6_C
Salsolinol	$C_{10}H_{13}NO_2$	179.0946	1.07	P6_C
Salicylamide	$C_7H_7NO_2$	137.0477	1.17	P6_C
3-hydroxybenzoic acid	$C_7H_6O_3$	138.0317	1.33	P6_D
Ketoleucine	$C_6H_{10}O_3$	130.0630	1.50	P6_D
3-alpha,11-beta,17-alpha,21-tetrahydroxy- 5-alpha-pregnan-20-one 3,21-diacetate	$C_{21}H_{34}O_5$	366.2406	1.11	P6_D
N,N-dimethyl-1,4-phenylenediamine	$C_8H_{12}N_2$	136.1000	1.25	P6_D
Homogentisic acid	$C_8H_8O_4$	168.0423	1.00	P6_D
Indoleacetaldehyde	$C_{10}H_9NO$	159.0684	1.55	P6_D
4-hydroxy-3-methoxyphenylglycol	$C_9H_{12}O_4$	184.0736	0.11	P6_D
3-hydroxyphenylacetic acid	$C_8H_8O_3$	152.0473	1.31	P6_D
4-methylcatechol	$C_7H_8O_2$	124.0524	1.88	P6_D
Pyridoxal	$C_8H_9NO_3$	167.0582	0.18	P6_D
Salicylic acid	$C_7H_6O_3$	138.0317	1.98	P6_D
Sebacic acid	$C_{10}H_{18}O_4$	202.1205	2.27	P6_D
3-methyl-2-oxindole	C_9H_9NO	147.0684	1.62	P6_E
3-methyladenine	$C_6H_7N_5$	149.0701	-0.31	P6_E
Hydroxyphenyllactic acid	$C_9H_{10}O_4$	182.0579	0.88	P6_E
Biotin	$C_{10}H_{16}N_2O_3S$	244.0882	0.32	P6_E
Mercaptopyruvic acid	$C_3H_4O_3S$	119.9881	0.29	P6_E

Name	formula	exact mass	logP	Mix
Pyruvic aldehyde	$C_3H_4O_2$	72.0211	0.20	P6_E
Pyrrole-2-carboxylic acid	$C_5H_5NO_2$	111.0320	0.63	P6_E
5-hydroxyindoleacetic acid	$C_{10}H_9NO_3$	191.0582	1.41	P6_E
Phenylacetic acid	$C_8H_8O_2$	136.0524	1.61	P6_E
Resorcinol monoacetate	$C_8H_8O_3$	152.0473	1.28	P6_E
Acetoacetic acid	$C_4H_6O_3$	102.0317	0.00	P6_E
Acetylphosphate	$C_2H_5O_5P$	139.9875	-0.88	P6_E
Ethyl 3-indoleacetic acid	$C_{12}H_{13}NO_2$	203.0946	2.21	P6_F
Dehydroascorbate	$C_6H_6O_6$	174.0164	-0.67	P6_F
Fructose	$C_6H_{12}O_6$	180.0634	-2.76	P6_F
Sorbose	$C_6H_{12}O_6$	180.0634	-2.76	P6_F
Xylitol	$C_5H_{12}O_5$	152.0685	-3.10	P6_F
Ribitol	$C_5H_{12}O_5$	152.0685	-3.10	P6_F
Myoinositol	$C_6H_{12}O_6$	180.0634	-3.78	P6_F
Mannose	$C_6H_{12}O_6$	180.0634	-2.93	P6_F
Arabinose	$C_5H_{10}O_5$	150.0528	-2.30	P6_F
Xylose	$C_5H_{10}O_5$	150.0528	-2.30	P6_G
Sucrose	$C_{12}H_{22}O_{11}$	342.1162	-4.53	P6_G
Galactose	$C_6H_{12}O_6$	180.0634	-2.93	P6_G
α-D-glucose	$C_6H_{12}O_6$	180.0634	-2.93	P6_G
Allose	$C_6H_{12}O_6$	180.0634	-2.93	P6_G
Mannitol	$C_6H_{14}O_6$	182.0790	-3.73	P6_G
Melibiose	$C_{12}H_{22}O_{11}$	342.1162	-4.70	P6_G

Name	formula	exact mass	logP	Mix
Sorbitol	$C_6H_{14}O_6$	182.0790	-3.73	P6_G
Maltose	$C_{12}H_{22}O_{11}$	342.1162	-4.70	P6_G
Tagatose	$C_6H_{12}O_6$	180.0634	-3.27	P6_G
L-gulonolactone	$C_6H_{10}O_6$	178.0477	-2.75	P6_G
Arabinose	$C_5H_{10}O_5$	150.0528	-2.94	P6_G
Cellobiose	$C_{12}H_{22}O_{11}$	342.1162	-4.70	P6_H
Psicose	$C_6H_{12}O_6$	180.0634	-2.76	P6_H
Arabitol	$C_5H_{12}O_5$	152.0685	-3.10	P6_H
Lyxose	$C_5H_{10}O_5$	150.0523	-2.61	P6_H
Ribose	$C_5H_{10}O_5$	150.0528	-2.30	P6_H
Palatinose	$C_{12}H_{22}O_{11}$	342.1162	-4.53	P6_H
Vitamin D2	$C_{28}H_{44}O$	396.3392	7.05	P7_A
Squalene	$C_{30}H_{50}$	410.3913	10.42	P7_A
4-coumaric acid	$C_9H_8O_3$	164.0473	1.83	P7_A
Nonanoic acid	$C_9H_{18}O_2$	158.1307	3.14	P7_A
Estradiol-17alpha	$C_{18}H_{24}O_2$	272.1776	3.75	P7_A
Caprylic acid	$C_8H_{16}O_2$	144.1150	2.70	P7_A
Ursodeoxycholic acid	$C_{24}H_{40}O_4$	392.2927	3.71	P7_A
Petroselinic acid	$C_{18}H_{34}O_2$	282.2559	6.78	P7_A
Dipalmitoylglycerol	$C_{35}H_{68}O_5$	568.5067	12.00	P7_A
Deoxycholic acid	$C_{24}H_{40}O_4$	392.2927	3.79	P7_A
Lithocholic acid	$C_{24}H_{40}O_3$	376.2977	5.02	P7_A
Protoporphyrin	$C_{34}H_{34}N_4O_4$	562.2580	6.58	P7_A

Name	formula	exact mass	logP	Mix
Heptanoic acid	$C_7H_{14}O_2$	130.0994	2.26	P7_B
Retinol	$C_{20}H_{30}O$	286.2297	4.69	P7_B
Menaquinone	$C_{41}H_{56}O_2$	580.4280	11.80	P7_B
Elaidic acid	$C_{18}H_{34}O_2$	282.2559	6.78	P7_B
Chenodeoxycholic acid	$C_{24}H_{40}O_4$	392.2927	3.71	P7_B
Myristic acid	$C_{14}H_{28}O_2$	228.2089	5.37	P7_B
Cholesteryl oleate	$C_{45}H_{78}O_2$	650.6002	14.56	P7_B
Rosmarinic acid	$C_{18}H_{16}O_8$	360.0845	3.00	P7_B
Glyceryl tripalmitate	$C_{51}H_{98}O_6$	806.7363	18.92	P7_B
Cortexolone	$C_{21}H_{30}O_4$	346.2144	2.58	P7_B
Lithocholyltaurine	$C_{26}H_{45}NO_5S$	483.3018	2.48	P7_B
Palmitoleic acid	$C_{16}H_{30}O_2$	254.2246	5.89	P7_B
Palmitic acid	$C_{16}H_{32}O_2$	256.2402	6.26	P7_C
Liothyronine	$C_{15}H_{12}I_3NO_4$	650.7900	2.80	P7_C
Sphinganine	$C_{18}H_{39}NO_2$	301.2981	4.77	P7_C
Lanosterol	$C_{30}H_{50}O$	426.3862	7.71	P7_C
Lauric acid	$C_{12}H_{24}O_2$	200.1776	4.48	P7_C
Arachidic acid	$C_{20}H_{40}O_2$	312.3028	8.03	P7_C
Erucic acid	$C_{22}H_{42}O_2$	338.3185	8.56	P7_C
Deoxycholic acid	$C_{24}H_{40}O_4$	392.2927	3.79	P7_C
Ketoleucine	$C_6H_{10}O_3$	130.0630	1.50	P7_C
Eicosapentaenoic acid	$C_{20}H_{30}O_2$	302.2246	6.23	P7_C
Heptadecanoic acid	$C_{17}H_{34}O_2$	270.2559	6.70	P7_C

Name	formula	exact mass	logP	Mix
Glyceryl trimyristate	$C_{45}H_{86}O_6$	722.6424	16.26	P7_C
Linoleic acid	$C_{18}H_{32}O_2$	280.2402	6.42	P7_D
Sphingomyelin	$C_{41}H_{83}N_2O_6P$	730.5989	7.87	P7_D
7-dehydrocholesterol	$C_{27}H_{44}O$	384.3392	6.71	P7_D
Thyroxine	$C_{15}H_{11}I_4NO_4$	776.6867	3.73	P7_D
Bis(2-ethylhexyl)phthalate	$C_{24}H_{38}O_4$	390.2770	8.03	P7_D
γ-linolenic acid	$C_{18}H_{30}O_2$	278.2246	6.06	P7_D
ω-hydroxydodecanoic acid	$C_{12}H_{24}O_3$	216.1725	3.04	P7_D
Methyl jasmonate	$C_{13}H_{20}O_3$	224.1412	2.56	P7_D
Dipalmitoyl-phosphatidylcholine	$C_{40}H_{80}NO_8P$	733.5622	8.11	P7_D
Hexadecanol	$C_{16}H_{34}O$	242.2610	6.14	P7_D
5,6 dimethylbenzimidazole	$C_9H_{10}N_2$	146.0844	2.29	P7_D
Retinoic acid	$C_{20}H_{28}O_2$	300.2089	5.01	P7_D
Indole	C_8H_7N	117.0578	2.07	P7_E
Cholic acid	$C_{24}H_{40}O_5$	408.2876	2.48	P7_E
Phylloquinone	$C_{31}H_{46}O2$	450.3498	9.70	P7_E
Cholesteryl palmitate	$C_{43}H_{76}O_2$	624.5845	14.03	P7_E
Quinoline	C_9H_7N	129.0578	2.13	P7_E
Docosahexaenoic acid	$C_{22}H_{32}O_2$	328.2402	6.75	P7_E
Diethyl 2-methyl-3-oxosuccinate	$C_9H_{14}O_5$	202.0841	1.74	P7_E
Retinyl palmitate	$C_{36}H_{60}O_2$	524.4593	11.62	P7_E
2-undecanone	$C_{11}H_{22}O$	170.1671	3.92	P7_E
1-hydroxy-2-naphthoate	$C_{11}H_8O_3$	188.0473	2.97	P7_E

Name	formula	exact mass	logP	Mix
Dipalmitoyl-phosphoethanolamine	$C_{37}H_{74}NO_8P$	691.5152	10.45	P7_E
Phenylpyruvic acid	$C_9H_8O_3$	164.0473	1.90	P7_E
Trans-cinnamic acid	$C_9H_8O_2$	148.0524	2.14	P7_F
Oleic acid	$C_{18}H_{34}O_2$	282.2559	6.78	P7_F
Stearic acid	$C_{18}H_{36}O_2$	284.2715	7.15	P7_F
β-carotene	$C_{40}H_{56}$	536.4382	11.12	P7_F
25-hydroxycholesterol	$C_{27}H_{46}O_2$	402.3498	5.64	P7_F
Nervonic acid	$C_{24}H_{46}O_2$	366.3498	9.45	P7_F
Desmosterol	$C_{27}H_{44}O$	384.3392	6.71	P7_F
Deoxycorticosterone acetate	$C_{23}H_{32}O_4$	372.2301	3.77	P7_F
Oleoyl-glycerol	$C_{21}H_{40}O_4$	356.2927	5.61	P7_F
α-tocopherol	$C_{29}H_{50}O_2$	430.3811	10.51	P7_F
Glycerol-myristate	$C_{17}H_{34}O_4$	302.2457	4.19	P7_F
Tricosanoic acid	$C_{23}H_{46}O_2$	354.3498	9.37	P7_F
Coenzyme Q10	$C_{59}H_{90}O_4$	862.6839	17.16	P7_G
Cortisone	$C_{21}H_{28}O_5$	360.1937	1.66	P7_G
Decanoic acid	$C_{10}H_{20}O_2$	172.1463	3.59	P7-G
Corticosterone	$C_{21}H_{30}O_4$	346.2144	2.02	P7_G

SI Table 2: Metabolites used in the interlaboratory study in chapter 3.

Name	formula	exact mass	logP	Mix
Glycine	$C_2H_5NO_2$	75.0320	-3.41	A
L-Tryptophan	$C_{11}H_{12}N_2O_2$	204.0899	-1.09	A
L-Valine	$C_5H_{11}NO_2$	117.0790	-1.95	A
L-Glutamine	$C_5H_{10}N_2O_3$	146.0691	-4	A
L-Proline	$C_5H_9NO_2$	115.0633	-2.57	A
Glutathione, reduced (GSH)	$C_{10}H_{17}N_3O_6S$	307.0838	-4.88	A
Palmitic acid	$C_{16}H_{32}O_2$	256.2402	6.26	A
Anthranilic acid	$C_7H_7NO_2$	137.0477	1.45	A
Homogentisic acid	$C_8H_8O_4$	168.0423	1	A
PC 16:0/0:0	$C_{24}H_{50}NO_7P$	495.3325	1.19	A
Benzoic acid	$C_7H_6O_2$	122.0368	1.63	A
Cholesterol	$C_{27}H_{46}O$	386.3549	7.11	A
Vitamin K1	$C_{31}H_{46}O_2$	450.3498	9.7	A
gamma-Linolenic acid	$C_{18}H_{30}O_2$	278.2246	6.06	A
Coenzyme Q2	$C_{19}H_{26}O_4$	318.1831	3.88	A
Kaempferol	$C_{15}H_{10}O_6$	286.0477	2.46	A
3-Hydroxyphenylacetic acid	$C_8H_8O_3$	152.0473	1.31	A
25-Hydroxyvitamin D2	$C_{28}H_{44}O_2$	412.3341	5.66	A
Ala-Phe	$C_{12}H_{16}N_2O_3$	236.1161	-1.72	A
N-Stearoyl Taurine	$C_{20}H_{41}NO_4S$	391.2756	4.64	A
L-Alanine	$C_3H_7NO_2$	89.0477	-2.84	B
L-Leucine	$C_6H_{13}NO_2$	131.0946	-1.59	B

Name	formula	exact mass	logP	Mix
L-Threonine	$C_4H_9NO_3$	119.0582	-3.47	B
L-Glutamic acid	$C_5H_9NO_4$	147.0532	-3.24	B
L-Cysteine	$C_3H_7NO_2S$	121.0197	-2.79	B
Glutathione, oxidized (GSSG)	$C_{20}H_{32}N_6O_{12}S_2$	612.1520	-10.11	B
Myristic acid	$C_{14}H_{28}O_2$	228.2089	5.37	B
2-Heptyl-3-hydroxy-4(1H)-quinolone	$C_{16}H_{21}NO_2$	259.1572	4.65	B
4-Hydroxybenzoic acid	$C_7H_6O_3$	138.0317	1.33	B
Pantothenic acid	$C_9H_{17}NO_5$	219.1107	-1.36	B
Xanthine	C5H4N4O2	152.0334	-0.21	B
3-Methylhippuric acid	$C_{10}H_{11}NO_3$	193.0739	1.04	B
Folic acid	$C_{19}H_{19}N_7O_6$	441.1397	-1.2	B
FAD	$C_{27}H_{33}N_9O_{15}P_2$	785.1571	-5.52	B
L-Cystine	$C_6H_{12}N_2O_4S_2$	240.0238	-5.9	B
Trehalose	$C_{12}H_{22}O_{11}$	342.1162	-4.7	B
Isovaleric acid	$C_5H_{10}O_2$	102.0681	1.21	B
2-Oxovaleric acid	$C_5H_8O_3$	116.0473	1.21	B
Stearoyl ethanolamide	$C_{20}H_{41}NO_2$	327.3137	5.87	B
Linoleyl ethanolamide	$C_{20}H_{37}NO_2$	323.2824	5.15	B
L-Tyrosine	$C_9H_{11}NO_3$	181.0739	-1.49	C
L-Serine	$C_3H_7NO_3$	105.0426	-3.89	C
L-Lysine	$C_6H_{14}N_2O_2$	146.1055	-3.21	C
L-Asparagine	$C_4H_8N_2O_3$	132.0535	-4.29	C
L-Methionine	$C_5H_{11}NO_2S$	149.0510	-2.19	C

Name	formula	exact mass	logP	Mix
Adenosine	$C_{10}H_{13}N_5O_4$	267.0968	-2.09	C
Ferulic acid	$C_{10}H_{10}O_4$	194.0579	1.67	C
4-Hydroxyphenylpyruvic acid	$C_9H_8O_4$	180.0423	1.6	C
Pyrocatechol	$C_6H_6O_2$	110.0368	1.37	C
Xanthosine	$C_{10}H_{12}N_4O_6$	284.0757	-1.81	C
Hippuric acid	$C_9H_9NO_3$	179.0582	0.53	C
Retinoic acid	$C_{20}H_{28}O_2$	300.2089	5.01	C
CoA	$C_{21}H_{36}N_7O_{16}P_3S$	767.1152	-5.77	C
N-Acetyl-D-Mannosamine	$C_8H_{15}NO_6$	221.0899	-3.86	C
Luteolin	$C_{15}H_{10}O_6$	286.0477	2.4	C
Suberic acid	$C_8H_{14}O_4$	174.0892	1.38	C
3-Methyladipic acid	$C_7H_{12}O_4$	160.0736	0.78	C
Lipoic acid, reduced	$C_8H_{16}O_2S_2$	208.0592	2.2	C
Pregnenolone sulfate	$C_{21}H_{32}O_2$	316.2402	3.58	C
L-Phenylalanine	$C_9H_{11}NO_2$	165.0790	-1.18	D
L-Isoleucine	$C_6H_{13}NO_2$	131.0946	-1.51	D
L-Arginine	$C_6H_{14}N_4O_2$	174.1117	-3.16	D
L-Aspartic acid	$C_4H_7NO_4$	133.0375	-3.5	D
L-Histidine	$C_6H_9N_3O_2$	155.0695	-3.62	D
Inosine	$C_{10}H_{12}N_4O_5$	268.0808	-1.57	D
Chorismic acid	$C_{10}H_{10}O_6$	226.0477	-0.13	D
Phenylpyruvic acid	$C_9H_8O_3$	164.0473	1.9	D
Arachidonic acid	$C_{20}H_{32}O_2$	304.2402	6.59	D

Name	formula	exact mass	logP	Mix
cAMP	$C_{10}H_{12}N_5O_6P$	329.0525	-4.02	D
Riboflavin	$C_{17}H_{20}N_4O_6$	376.1383	-0.92	D
Uric acid	$C_5H_4N_4O_3$	168.0283	0.75	D
Acetyl-CoA	$C_{23}H_{38}N_7O_{17}P_3S$	809.1258	-5.92	D
L-Cystathionine	$C_7H_{14}N_2O_4S$	222.0674	-5.82	D
Retinol	$C_{20}H_{30}O$	286.2297	4.69	D
4-Methylvaleric acid	$C_6H_{12}O_2$	116.0837	1.65	D
α-Hydroxyhippuric acid	$C_9H_9NO_4$	195.0532	0.24	D
7-Methylxanthine	$C_6H_6N_4O_2$	166.0491	0.02	D
Indolacetic acid	$C_{10}H_9NO_2$	175.0633	1.71	D
L-tert-Leucine	$C_6H_{13}NO_2$	131.0946	-1.57	E
L-Norleucine	$C_6H_{13}NO_2$	131.0946	-1.43	E
Theobromine	$C_7H_8N_4O_2$	180.0647	-0.77	E
Quercetin	$C_{15}H_{10}O_7$	302.0427	2.16	E
Decanoyl-L-carnitine	$C_{17}H_{34}NO_4$	316.2482	-0.63	E
Palmitoyl-L-carnitine	$C_{23}H_{46}NO_4$	400.3421	2.03	E
Oleoyl-L-carnitine	$C_{25}H_{48}NO_4$	426.3578	2.56	E
Acetyl-L-carnitine	$C_9H_{17}NO_4$	203.1158	-4.45	E
Octanoyl-L-carnitine	$C_{15}H_{30}NO_4$	288.2169	-1.52	E
Tetradecanoyl-L-carnitine	$C_{21}H_{42}NO_4$	372.3108	1.14	E
Hexanoyl-L-carnitine	$C_{13}H_{26}NO_4$	260.1856	-2.41	E
Stearoyl-L-Carnitine	$C_{25}H_{50}NO_4$	428.3734	2.92	E
Reserpine	$C_{33}H_{40}N_2O_9$	608.2734	3.53	E

Name	formula	exact mass	logP	Mix
Orotic acid	$C_5H_4N_2O_4$	156.0171	-1.23	E
all trans-Retinal	$C_{20}H_{28}O$	284.2140	4.86	E
D-Sphingosine	$C_{18}H_{37}NO_2$	299.2824	4.57	E
Apigenin	$C_{15}H_{10}O_5$	270.0528	2.71	E
2-Hydroxyphenylacetic acid	$C_8H_8O_3$	152.0473	1.31	E
N-α-Acetyl-L-ornithine	$C_7H_{14}N_2O_3$	174.1004	-3.6	E
Leu-Gly-Gly	$C_{10}H_{19}N_3O_4$	245.1376	-3.8	E
Acetaminophen	$C_8H_9NO_2$	151.0633	0.91	E

8.2.2. Lipids

SI Table 3: Ordering information of lipids obtained from Avanti Polar Lipids

Name	Ordering Number	Formula
PC 14:0/14:0	850345C-25MG	$C_{36}H_{72}NO_8P$
PC 15:0/15:0	850350C-25MG	$C_{38}H_{76}NO_8P$
PC 16:0/16:0	850355C-25MG	$C_{40}H_{80}NO_8P$
PC 16:0/18:0	850456C-25MG	$C_{42}H_{84}NO_8P$
PC 16:0/18:1(9Z)	850457C-25MG	$C_{42}H_{82}NO_8P$
PC 16:0/18:2(9Z,12Z)	850458C-25MG	$C_{42}H_{80}NO_8P$
PC 16:0/20:4(5Z,8Z,11Z,14Z)	850459C-25MG	$C_{44}H_{80}NO_8P$
PC 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	850461C-25MG	$C_{46}H_{80}NO_8P$
PC 17:0/17:0	850360C-25MG	$C_{42}H_{84}NO_8P$
PC 18:0/16:0	850465C-25MG	$C_{42}H_{84}NO_8P$
PC 18:0/18:0	850365C-25MG	$C_{44}H_{88}NO_8P$

Name	Ordering Number	Formula
PC 18:0/18:1(9Z)	850467C-25MG	$C_{44}H_{86}NO_8P$
PC 18:0/18:2(9Z,12Z)	850468C-25MG	$C_{44}H_{84}NO_8P$
PC 18:0/20:4(5Z,8Z,11Z,14Z)	850469C-25MG	$C_{46}H_{84}NO_8P$
PC 18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	850472C-25MG	$C_{48}H_{84}NO_8P$
PC 18:1(9Z)/16:0	850475C-25MG	$C_{42}H_{82}NO_8P$
PC 18:1(9Z)/18:0	850476C-25MG	$C_{44}H_{86}NO_8P$
PC 20:0/20:0	850368C-25MG	$C_{48}H_{96}NO_8P$
PC 22:0/22:0	850371C-25MG	$C_{52}H_{104}NO_8P$
PE 14:0/14:0	850745P-25MG	$C_{33}H_{66}NO_8P$
PE 15:0/15:0	850704P-25MG	$C_{35}H_{70}NO_8P$
PE 16:0/16:0	850705P-25MG	$C_{37}H_{74}NO_8P$
PE 16:0/18:1(9Z)	850757C-25MG	$C_{39}H_{76}NO_8P$
PE 16:0/18:2(9Z,12Z)	850756C-25MG	$C_{39}H_{74}NO_8P$
PE 16:0/20:4(5Z,8Z,11Z,14Z)	850759C-25MG	$C_{41}H_{74}NO_8P$
PE 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	850801C-25MG	$C_{43}H_{74}NO_8P$
PE 17:0/17:0	830756P-25MG	$C_{39}H_{78}NO_8P$
PE 18:0/18:0	850715P-25MG	$C_{41}H_{82}NO_8P$
PE 18:0/18:1(9Z)	850758C-25MG	$C_{41}H_{80}NO_8P$
PE 18:0/18:2(9Z,12Z)	850802C-25MG	$C_{41}H_{78}NO_8P$
PE 18:0/20:4(5Z,8Z,11Z,14Z)	850804C-25MG	$C_{43}H_{78}NO_8P$
PE 18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	850806C-25MG	$C_{45}H_{78}NO_8P$
PS 14:0/14:0	840033P-10MG	$C_{34}H_{66}NO_{10}P$
PS 16:0/16:0	840037P-10MG	$C_{38}H_{74}NO_{10}P$

Name	Ordering Number	Formula
PS 16:0/18:1(9Z)	840034C-10MG	$C_{40}H_{76}NO_{10}P$
PS 16:0/18:2(9Z,12Z)	840060C-10MG	$C_{40}H_{74}NO_{10}P$
PS 16:0/20:4(5Z,8Z,11Z,14Z)	840061C-10MG	$C_{42}H_{74}NO_{10}P$
PS 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	840062C-25MG	$C_{44}H_{74}NO_{10}P$
PS 17:0/17:0	840028P-10MG	$C_{40}H_{78}NO_{10}P$
PS 18:0/18:0	840029P-10MG	$C_{42}H_{82}NO_{10}P$
PA 14:0/14:0	830845P-25MG	$C_{31}H_{61}O_8P$
PA 16:0/16:0	830855P-25MG	$C_{35}H_{69}O_8P$
PA 16:0/18:1(9Z)	840857C-25MG	$C_{37}H_{71}O_8P$
PA 16:0/18:2(9Z,12Z)	840858C-25MG	$C_{37}H_{69}O_8P$
PA 16:0/20:4(5Z,8Z,11Z,14Z)	840859C-25MG	$C_{39}H_{69}O_8P$
PA 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	840860C-25MG	$C_{41}H_{69}O_8P$
PA 17:0/17:0	830856P-25MG	$C_{37}H_{73}O_8P$
PA 18:0/18:0	830865P-25MG	$C_{39}H_{77}O_8P$
PG 14:0/14:0	840445P-25MG	$C_{34}H_{67}O_{10}P$
PG 15:0/15:0	840446P-25MG	$C_{36}H_{71}O_{10}P$
PG 16:0/16:0	840455P-25MG	$C_{38}H_{75}O_{10}P$
PG 16:0/18:1(9Z)	840457C-25MG	$C_{40}H_{77}O_{10}P$
PG 16:0/18:2(9Z,12Z)	840497C-25MG	$C_{40}H_{75}O_{10}P$
PG 16:0/20:4(5Z,8Z,11Z,14Z)	840499C-25MG	$C_{42}H_{75}O_{10}P$
PG 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	840500C-25MG	$C_{44}H_{75}O_{10}P$
PG 17:0/17:0	830456P-25MG	$C_{40}H_{79}O_{10}P$
PG 18:0/18:0	840465P-25MG	$C_{42}H_{83}O_{10}P$

8.2.3. CCS values

SI Table 4: CCS values of lipids obtained in positive ionization mode

Shorthand	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield	$^{DT}CCS_{N2}$ Single Field 1	$^{DT}CCS_{N2}$ Single Field 2
PC 14:0/14:0	$[M+Na]^+$	700.4888	273.30	NA	276.33
PC 15:0/15:0	$[M+Na]^+$	728.5201	279.00	NA	281.50
PC 16:0/16:0	$[M+Na]^+$	756.5514	284.30	NA	287.43
PC 16:0/18:0	$[M+Na]^+$	784.5827	290.90	295.30	293.97
PC 16:0/18:1(9Z)	$[M+Na]^+$	782.5670	289.53	293.20	292.60
PC 16:0/18:2(9Z,12Z)	$[M+Na]^+$	780.5514	287.40	291.00	290.60
PC 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+Na]^+$	804.5514	290.70	294.40	292.90
PC 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+Na]^+$	828.5514	293.00	297.20	295.47
PC 17:0/17:0	$[M+Na]^+$	784.5827	290.93	294.80	293.40
PC 18:0/16:0	$[M+Na]^+$	784.5827	291.47	294.90	293.50
PC 18:0/18:0	$[M+Na]^+$	812.6140	296.03	300.80	298.90
PC 18:0/18:1(9Z)	$[M+Na]^+$	810.5983	295.40	299.00	298.30
PC 18:0/18:2(9Z,12Z)	$[M+Na]^+$	808.5827	293.40	297.00	296.83
PC 18:0/20:4(5Z,8Z,11Z,14Z)	$[M+Na]^+$	832.5827	296.40	300.30	300.37
PC 18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+Na]^+$	856.5827	299.00	299.70	303.47
PC 18:1(9Z)/16:0	$[M+Na]^+$	782.5670	290.43	293.90	294.47
PC 18:1(9Z)/18:0	$[M+Na]^+$	810.5983	296.23	299.80	NA
PC 20:0/20:0	$[M+Na]^+$	868.6766	309.73	313.70	NA
PC 22:0/22:0	$[M+Na]^+$	924.7392	321.45	NA	323.90
PE 14:0/14:0	$[M+H]^+$	636.4599	260.11	262.30	NA
PE 15:0/15:0	$[M+H]^+$	664.4912	266.46	267.40	NA

Shorthand	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield	$^{DT}CCS_{N2}$ Single Field 1	$^{DT}CCS_{N2}$ Single Field 2
PE 16:0/16:0	$[M+H]^+$	692.5225	272.45	273.10	NA
PE 16:0/18:1(9Z)	$[M+H]^+$	718.5381	275.30	277.00	NA
PE 16:0/18:2(9Z,12Z)	$[M+H]^+$	716.5225	273.21	274.60	NA
PE 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+H]^+$	740.5225	276.90	278.30	NA
PE 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+H]^+$	764.5225	279.19	280.80	NA
PE 17:0/17:0	$[M+H]^+$	720.5538	278.47	279.20	NA
PE 18:0/18:0	$[M+H]^+$	748.5851	284.45	285.40	NA
PE 18:0/18:1(9Z)	$[M+H]^+$	746.5694	281.65	283.60	NA
PE 18:0/18:2(9Z,12Z)	$[M+H]^+$	744.5538	279.53	281.10	NA
PE 18:0/20:4(5Z,8Z,11Z,14Z)	$[M+H]^+$	768.5538	283.26	284.90	NA
PE 18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+H]^+$	792.5538	285.72	287.40	NA
PE 14:0/14:0	$[M+Na]^+$	658.4418	263.43	266.20	NA
PE 15:0/15:0	$[M+Na]^+$	686.4731	269.37	271.20	NA
PE 16:0/16:0	$[M+Na]^+$	714.5044	275.47	NA	NA
PE 16:0/18:1(9Z)	$[M+Na]^+$	740.5201	281.03	283.10	NA
PE 16:0/18:2(9Z,12Z)	$[M+Na]^+$	738.5044	277.67	279.70	NA
PE 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+Na]^+$	762.5044	282.53	284.20	NA
PE 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+Na]^+$	786.5044	285.23	287.10	NA
PE 17:0/17:0	$[M+Na]^+$	742.5357	281.93	NA	NA
PE 18:0/18:0	$[M+Na]^+$	770.5670	288.60	NA	NA
PE 18:0/18:1(9Z)	$[M+Na]^+$	768.5514	287.23	289.80	NA
PE 18:0/18:2(9Z,12Z)	$[M+Na]^+$	766.5357	284.50	286.20	NA

Shorthand	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield	$^{DT}CCS_{N2}$ Single Field 1	$^{DT}CCS_{N2}$ Single Field 2
PE 18:0/20:4(5Z,8Z,11Z,14Z)	$[M+Na]^+$	790.5357	288.73	290.60	NA
PE 18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+Na]^+$	814.5357	291.57	293.90	NA
PS 14:0/14:0	$[M+H]^+$	680.4497	268.12	269.40	NA
PS 16:0/16:0	$[M+H]^+$	736.5123	NA	281.10	NA
PS 16:0/18:1(9Z)	$[M+H]^+$	762.5280	281.37	283.10	NA
PS 16:0/18:2(9Z,12Z)	$[M+H]^+$	760.5123	NA	NA	NA
PS 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+H]^+$	784.5123	281.95	283.50	NA
PS 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+H]^+$	808.5123	284.71	285.70	NA
PS 17:0/17:0	$[M+H]^+$	764.5436	285.95	287.20	NA
PS 18:0/18:0	$[M+H]^+$	792.5749	NA	292.70	NA
PS 14:0/14:0	$[M+Na]^+$	702.4317	NA	269.20	NA
PS 16:0/16:0	$[M+Na]^+$	758.4943	NA	281.40	NA
PS 16:0/18:1(9Z)	$[M+Na]^+$	784.5099	NA	283.70	NA
PS 16:0/18:2(9Z,12Z)	$[M+Na]^+$	782.4943	NA	NA	NA
PS 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+Na]^+$	806.4943	NA	287.40	NA
PS 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+Na]^+$	830.4943	NA	289.90	NA
PS 17:0/17:0	$[M+Na]^+$	786.5256	NA	288.40	NA
PS 18:0/18:0	$[M+Na]^+$	814.5569	NA	294.60	NA
PG 14:0/14:0	$[M+H]^+$	667.4545	NA	275.40	NA
PG 15:0/15:0	$[M+H]^+$	695.4858	NA	281.20	NA
PG 16:0/16:0	$[M+H]^+$	723.5171	NA	286.90	NA
PG 16:0/18:1(9Z)	$[M+H]^+$	749.5327	NA	289.10	NA

Shorthand	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield	$^{DT}CCS_{N2}$ Single Field 1	$^{DT}CCS_{N2}$ Single Field 2
PG 16:0/18:2(9Z,12Z)	$[M+H]^+$	747.5171	NA	285.90	NA
PG 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+H]^+$	771.5171	NA	289.40	NA
PG 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+H]^+$	795.5171	NA	292.00	NA
PG 14:0/14:0	$[M+Na]^+$	689.4364	NA	270.40	NA
PG 15:0/15:0	$[M+Na]^+$	717.4677	NA	276.90	NA
PG 16:0/16:0	$[M+Na]^+$	745.4990	NA	283.10	NA
PG 16:0/18:1(9Z)	$[M+Na]^+$	771.5147	NA	284.10	NA
PG 16:0/18:2(9Z,12Z)	$[M+Na]^+$	769.4990	NA	281.90	NA
PG 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+Na]^+$	793.4990	NA	286.20	NA
PG 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+Na]^+$	817.4990	NA	289.00	NA
PG 17:0/17:0	$[M+Na]^+$	773.5303	NA	288.90	NA
PG 18:0/18:0	$[M+Na]^+$	801.5616	NA	294.70	NA

SI Table 5: CCS values of lipids obtained in negative ionization mode

Shorthand	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield	$^{DT}CCS_{N2}$ Single Field 1	$^{DT}CCS_{N2}$ Single Field 2
PC 14:0/14:0	$[M+FA-H]^-$	722.4978	275.34	275.20	NA
PC 15:0/15:0	$[M+FA-H]^-$	750.5291	281.30	280.30	NA
PC 16:0/16:0	$[M+FA-H]^-$	778.5604	286.67	285.30	NA
PC 16:0/18:0	$[M+FA-H]^-$	806.5917	291.95	290.70	NA
PC 16:0/18:1(9Z)	$[M+FA-H]^-$	804.5760	290.59	289.20	NA
PC 16:0/18:2(9Z,12Z)	$[M+FA-H]^-$	802.5604	289.66	287.90	NA
PC 16:0/20:4(5Z,8Z,11Z,14Z)	$[M+FA-H]^-$	826.5604	292.76	291.10	NA
PC 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+FA-H]^-$	850.5604	296.25	294.60	NA
PC 17:0/17:0	$[M+FA-H]^-$	806.5917	291.71	290.40	NA
PC 18:0/16:0	$[M+FA-H]^-$	806.5917	292.02	290.50	NA
PC 18:0/18:0	$[M+FA-H]^-$	834.6230	296.99	295.90	NA
PC 18:0/18:1(9Z)	$[M+FA-H]^-$	832.6073	295.77	294.60	NA
PC 18:0/18:2(9Z,12Z)	$[M+FA-H]^-$	830.5917	295.20	293.40	NA
PC 18:0/20:4(5Z,8Z,11Z,14Z)	$[M+FA-H]^-$	854.5917	298.00	296.40	NA
PC 18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M+FA-H]^-$	878.5917	301.43	300.10	NA
PC 18:1(9Z)/16:0	$[M+FA-H]^-$	804.5760	290.69	289.10	NA
PC 18:1(9Z)/18:0	$[M+FA-H]^-$	832.6073	296.13	294.70	NA
PC 20:0/20:0	$[M+FA-H]^-$	890.6856	308.18	306.90	NA
PC 22:0/22:0	$[M+FA-H]^-$	946.7482	319.06	NA	NA
PE 14:0/14:0	$[M-H]^-$	634.4453	249.17	249.30	NA
PE 15:0/15:0	$[M-H]^-$	662.4766	255.55	255.20	NA
PE 16:0/16:0	$[M-H]^-$	690.5079	262.46	262.10	NA

Shorthand	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield	$^{DT}CCS_{N2}$ Single Field 1	$^{DT}CCS_{N2}$ Single Field 2
PE 16:0/18:1(9Z)	$[M-H]^-$	716.5236	267.37	266.50	NA
PE 16:0/18:2(9Z,12Z)	$[M-H]^-$	714.5079	265.91	265.00	NA
PE 16:0/20:4(5Z,8Z,11Z,14Z)	$[M-H]^-$	738.5079	270.00	269.10	NA
PE 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M-H]^-$	762.5079	274.25	273.40	NA
PE 17:0/17:0	$[M-H]^-$	718.5392	268.70	269.20	NA
PE 18:0/18:0	$[M-H]^-$	746.5705	275.09	NA	NA
PE 18:0/18:1(9Z)	$[M-H]^-$	744.5549	273.76	273.30	NA
PE 18:0/18:2(9Z,12Z)	$[M-H]^-$	742.5392	272.39	271.90	NA
PE 18:0/20:4(5Z,8Z,11Z,14Z)	$[M-H]^-$	766.5392	276.11	275.70	NA
PE 18:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M-H]^-$	790.5392	280.53	280.00	NA
PS 14:0/14:0	$[M-H]^-$	678.4352	260.35	260.00	NA
PS 16:0/16:0	$[M-H]^-$	734.4978	272.14	273.10	NA
PS 16:0/18:1(9Z)	$[M-H]^-$	760.5134	276.94	276.30	NA
PS 16:0/18:2(9Z,12Z)	$[M-H]^-$	758.4978	#DIV/0!	NA	NA
PS 16:0/20:4(5Z,8Z,11Z,14Z)	$[M-H]^-$	782.4978	279.21	278.70	NA
PS 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M-H]^-$	806.4978	283.03	283.10	NA
PS 17:0/17:0	$[M-H]^-$	762.5291	278.14	278.80	NA
PS 18:0/18:0	$[M-H]^-$	790.5604	283.90	284.50	NA
PG 14:0/14:0	$[M-H]^-$	665.4399	256.60	257.20	NA
PG 15:0/15:0	$[M-H]^-$	693.4712	262.45	262.30	NA
PG 16:0/16:0	$[M-H]^-$	721.5025	269.11	268.30	NA
PG 16:0/18:1(9Z)	$[M-H]^-$	747.5182	274.04	273.20	NA

Shorthand	Adduct	*m/z*	$^{DT}CCS_{N2}$ Multifield	$^{DT}CCS_{N2}$ Single Field 1	$^{DT}CCS_{N2}$ Single Field 2
PG 16:0/18:2(9Z,12Z)	$[M-H]^-$	745.5025	272.34	271.40	NA
PG 16:0/20:4(5Z,8Z,11Z,14Z)	$[M-H]^-$	769.5025	276.20	275.50	NA
PG 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M-H]^-$	793.5025	280.31	279.80	NA
PG 17:0/17:0	$[M-H]^-$	749.5338	275.19	274.50	NA
PG 18:0/18:0	$[M-H]^-$	777.5651	281.15	281.00	NA
PA 14:0/14:0	$[M-H]^-$	591.4031	241.43	NA	NA
PA 16:0/16:0	$[M-H]^-$	647.4657	254.35	NA	NA
PA 16:0/18:1(9Z)	$[M-H]^-$	673.4814	259.45	NA	NA
PA 16:0/18:2(9Z,12Z)	$[M-H]^-$	671.4657	257.91	NA	NA
PA 16:0/20:4(5Z,8Z,11Z,14Z)	$[M-H]^-$	695.4657	262.22	NA	NA
PA 16:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)	$[M-H]^-$	719.4657	265.97	NA	NA
PA 17:0/17:0	$[M-H]^-$	675.4970	261.56	NA	NA
PA 18:0/18:0	$[M-H]^-$	703.5283	267.81	NA	NA

SI Table 6: CCS values of maradolipids

Ionization mode	Shorthand	Species	Adduct	RT [min]	$^{DT}CCS_{N2}$ [Å²]	m/z
Positive	Mar 14:0/14:0	Mar 28:0	$[M+NH_4]^+$	12.98	284.00	780.5484
	Mar 15:0/15:0	Mar 30:0	$[M+NH_4]^+$	13.90	290.60	808.5795
	Mar 14:0/18:1	Mar 32:1	$[M+NH_4]^+$	14.38	294.20	834.5949
	Mar 16:0/16:0	Mar 32:0	$[M+NH_4]^+$	15.10	296.57	836.6108
	Mar 15:0/17:0	Mar 32:0	$[M+NH_4]^+$	15.47	296.63	836.6108
	Mar 15:0/18:1	Mar 33:1	$[M+NH_4]^+$	14.78	297.37	848.6106
	Mar 16:0/18:1	Mar 34:1	$[M+NH_4]^+$	15.52	300.70	862.6264
	Mar 17:0/18:1	Mar 35:1	$[M+NH_4]^+$	15.85	303.17	876.6417
	Mar 18:1/18:1	Mar 36:2	$[M+NH_4]^+$	15.56	304.20	888.6419
	Mar 18:1/19:1	Mar 37:2	$[M+NH_4]^+$	16.23	307.67	902.6578
	Mar 14:0/14:0	Mar 28:0	$[M+Na]^+$	12.98	282.70	785.5032
	Mar 15:0/15:0	Mar 30:0	$[M+Na]^+$	13.90	289.43	813.5350
	Mar 14:0/18:1	Mar 32:1	$[M+Na]^+$	14.38	291.47	839.5501
	Mar 16:0/16:0	Mar 32:0	$[M+Na]^+$	15.09	295.30	841.5663
	Mar 15:0/17:0	Mar 32:0	$[M+Na]^+$	15.47	295.37	841.5660
	Mar 15:0/18:1	Mar 33:1	$[M+Na]^+$	14.78	294.93	853.5663
	Mar 16:0/18:1	Mar 34:1	$[M+Na]^+$	15.52	298.83	867.5820
	Mar 17:0/18:1	Mar 35:1	$[M+Na]^+$	15.85	301.73	881.5965
	Mar 18:1/18:1	Mar 36:2	$[M+Na]^+$	15.56	303.33	893.5975
	Mar 18:1/19:1	Mar 37:2	$[M+Na]^+$	16.23	306.70	907.6129
negative	Mar 14:0/14:0	Mar 28:0	$[M+FA-H]^-$	12.94	284.40	807.5119
	Mar 15:0/15:0	Mar 30:0	$[M+FA-H]^-$	13.87	290.77	835.5430

Mar 14:0/18:1	Mar 32:1	$[M+FA-H]^-$	14.35	294.37	861.5591
Mar 16:0/16:0	Mar 32:0	$[M+FA-H]^-$	15.06	297.03	863.5748
Mar 15:0/17:0	Mar 32:0	$[M+FA-H]^-$	15.44	297.07	863.5750
Mar 15:0/18:1	Mar 33:1	$[M+FA-H]^-$	14.74	297.10	875.5748
Mar 16:0/18:1	Mar 34:1	$[M+FA-H]^-$	15.48	300.07	889.5905
Mar 17:0/18:1	Mar 35:1	$[M+FA-H]^-$	15.82	303.17	903.6063
Mar 18:1/18:1	Mar 36:2	$[M+FA-H]^-$	15.53	303.13	915.6059
Mar 18:1/19:1	Mar 37:2	$[M+FA-H]^-$	16.20	307.17	929.6219

8.2.4. Building blocks in lipidomicsUtils and wormLipidBlastR

SI Table 7 building blocks of lipids

Name	Formula	charge	Mass / *m/z*
CoA (coa_mass)	$C_{21}H_{36}N_7O_{16}P_3S$	0	767.115210
Water (water_mass)	H_2O	0	18.010565
Phosphoric acid (h3po4_mass)	H_3PO_4	0	97.976896
Sulfuric acid (h2so4_mass)	H_2SO_4	0	97.967380
Hexose (hexose_mass)	$C_6H_{12}O_6$	0	180.063388
Di-hexose (dihexose_mass)	$C_{12}H_{22}O_{11}$	0	342.116212
Proton (proton_mass)	H^+	1+	1.007276
Sodium ion (sodium_ion_mass)	Na^+	1+	22.989221
Glycerophosphocholine (gpc_mass)	$C_8H_{20}NO_6P$	0	257.102824
Phosphochline (pc_mass)	$C_5H_{14}NO_4P$	0	183.066045
Choline (choline_mass)	$C_5H_{13}NO$	0	103.099714
Glycerophosphoethanolamine (gpe_mass)	$C_5H_{14}NO_6P$	0	215.055874

Name	Formula	charge	Mass / *m/z*
Phosphoethanolamine (pe_mass)	$C_2H_8NO_4P$	0	141.019095
Ethanolamine (ethanolamine_mass)	C_2H_7NO	0	61.052764
Glycerophosphoserine (gps_mass)	$C_6H_{14}NO_8P$	0	259.045703
Phosphoserine (ps_mass)	$C_3H_8NO_6P$	0	185.008924
Serine (serine_mass)	$C_3H_7NO_3$	0	105.042593
Glycerophosphoglycerol (gpg_mass)	$C_6H_{15}O_8P$	0	246.050454
Phosphoglycerol (pg_mass)	$C_3H_9O_6P$	0	172.013675
Glycerol	$C_3H_8O_3$	0	92.047344
Glycerophosphoinositol (gpi_mass)	$C_9H_{19}O_{11}P$	0	334.066498
Phosphoinositol (pi_mass)	$C_6H_{13}O_9P$	0	260.029719
Inositol (inositol_mass)	$C_6H_{12}O_6$	0	180.063388
Glycerophosphoinositolphosphate (gpip_mass)	$C_9H_{20}O_{14}P_2$	0	414.032829
Phosphoinositolphosphate (pip_mass)	$C_6H_{14}O_{12}P_2$	0	339.996050
Glycerophosphoinositol-di-phosphate (gpipp_mass)	$C_9H_{21}O_{17}P_3$	0	493.999160
Phosphoinositol-di-phosphate (pipp_mass)	$C_6H_{15}O_{15}P_3$	0	419.962381
Cardiolipin (cl_mass)	$C_9H_{22}O_{13}P_2$		400.053565

SI Table 8 Lipid Class specific fragments positive ion mode

Lipid class	Adduct	Fragment *m/z* / NL
PC	$[M+H]^+$	*m/z* 184.0733 NL 183.0661
PE	$[M+H]^+$	NL 141.0191
MMPE	$[M+H]^+$	NL 155.0347
DMPE	$[M+H]^+$	NL 169.0503 *m/z* 170.0576
PS	$[M+H]^+$	NL 185.0089
PG	$[M+NH_4]^+$	NL 189.0402
PA	$[M+NH_4]^+$	NL 115.0035 NL 17.0266
PI	$[M+NH_4]^+$	NL 277.0563
DG	$[M+NH_4]^+$	NL 35.0371 NL 17.0266
SM	$[M+H]^+$	*m/z* 184.0733
Cer	$[M+H]^+$	*m/z* 250.2529 *m/z* 268.2635 *m/z* 238.2530

SI Table 9 Lipid Class specific fragments negative ion mode

Lipid class	Adduct	Fragment m/z / NL
PC	$[M+FA-H]^-$	NL 60.0211
PE	$[M-H]^-$	*m/z* 140.0118 *m/z* 196.0380
PG	$[M-H]^-$	*m/z* 152.9958 *m/z* 171.0063
PA	$[M-H]^-$	*m/z* 152.9958
PI	$[M-H]^-$	*m/z* 241.0019 *m/z* 223.0013 *m/z* 152.9958 *m/z* 259.0224
PS	$[M-H]^-$	*m/z* 152.9958 NL 87.0320
DMPE	$[M-H]^-$	*m/z* 168.0431
SM	$[M+FA-H]^-$	*m/z* 168.0431 NL 60.0211

SI Table 10: Fatty acids for prediction and search

species level	Formula	exact mass	PI, [M-H]-	NL, NH4OOCR
FA 12:0	$C_{12}H_{24}O_2$	200.17763	199.170354	217.204179
FA 12:0;OH	$C_{12}H_{24}O_3$	216.172545	215.165269	233.199094
FA 13:0	$C_{13}H_{26}O_2$	214.19328	213.186004	231.219829
FA 13:0;OH	$C_{13}H_{26}O_3$	230.188195	229.180919	247.214744
FA 14:0	$C_{14}H_{28}O_2$	228.20893	227.201654	245.235479
FA 14:0;OH	$C_{14}H_{28}O_3$	244.203845	243.196569	261.230394
FA 15:0	$C_{15}H_{30}O_2$	242.22458	241.217304	259.251129
FA 15:0;OH	$C_{15}H_{30}O_3$	258.219495	257.212219	275.246044
FA 16:0	$C_{16}H_{32}O_2$	256.24023	255.232954	273.266779
FA 16:0;OH	$C_{16}H_{32}O_3$	272.235145	271.227869	289.261694
FA 16:1	$C_{16}H_{30}O_2$	254.22458	253.217304	271.251129
FA 17:0	$C_{17}H_{34}O_2$	270.25588	269.248604	287.282429
FA 17:0;OH	$C_{17}H_{34}O_3$	286.250795	285.243519	303.277344
FA 17:1	$C_{17}H_{32}O_2$	268.24023	267.232954	285.266779
FA 18:0	$C_{18}H_{36}O_2$	284.27153	283.264254	301.298079
FA 18:0;OH	$C_{18}H_{36}O_3$	300.266445	299.259169	317.292994
FA 18:1	$C_{18}H_{34}O_2$	282.25588	281.248604	299.282429
FA 18:2	$C_{18}H_{32}O_2$	280.24023	279.232954	297.266779
FA 18:3	$C_{18}H_{30}O_2$	278.22458	277.217304	295.251129
FA 18:4	$C_{18}H_{28}O_2$	276.20893	275.201654	293.235479
FA 19:0	$C_{19}H_{38}O_2$	298.28718	297.279904	315.313729
FA 19:0;OH	$C_{19}H_{38}O_3$	314.282095	313.274819	331.308644

species level	Formula	exact mass	PI, [M-H]-	NL, NH4OOCR
FA 19:1	$C_{19}H_{36}O_2$	296.27153	295.264254	313.298079
FA 20:0	$C_{20}H_{40}O_2$	312.302831	311.295555	329.32938
FA 20:0;OH	$C_{20}H_{40}O_3$	328.297745	327.290469	345.324294
FA 20:1	$C_{20}H_{38}O_2$	310.28718	309.279904	327.313729
FA 20:2	$C_{20}H_{36}O_2$	308.27153	307.264254	325.298079
FA 20:3	$C_{20}H_{34}O_2$	306.25588	305.248604	323.282429
FA 20:4	$C_{20}H_{32}O_2$	304.24023	303.232954	321.266779
FA 20:5	$C_{20}H_{30}O_2$	302.22458	301.217304	319.251129
FA 21:0	$C_{21}H_{42}O_2$	326.318481	325.311205	343.34503
FA 21:0;OH	$C_{21}H_{42}O_3$	342.313395	341.306119	359.339944
FA 22:0	$C_{22}H_{44}O_2$	340.334131	339.326855	357.36068
FA 22:0;OH	$C_{22}H_{44}O_3$	356.329045	355.321769	373.355594
FA 23:0	$C_{23}H_{46}O_2$	354.349781	353.342505	371.37633
FA 23:0;OH	$C_{23}H_{46}O_3$	370.344695	369.337419	387.371244
FA 24:0	$C_{24}H_{48}O_2$	368.365431	367.358155	385.39198
FA 24:0;OH	$C_{24}H_{48}O_3$	384.360345	383.353069	401.386894
FA 25:0	$C_{25}H_{50}O_2$	382.381081	381.373805	399.40763
FA 25:0;OH	$C_{25}H_{50}O_3$	398.375995	397.368719	415.402544
FA 26:0;	$C_{26}H_{52}O_2$	396.396731	395.389455	413.42328
FA 26:0;OH	$C_{26}H_{52}O_3$	412.391646	411.38437	429.418195
FA 27:0	$C_{27}H_{54}O_2$	410.412381	409.405105	427.43893
FA 27:0;OH	$C_{27}H_{54}O_3$	426.407296	425.40002	443.433845
FA 28:0	$C_{28}H_{56}O_2$	424.428031	423.420755	441.45458

species level	Formula	exact mass	PI, [M-H]-	NL, NH4OOCR
FA 28:0;OH	$C_{28}H_{56}O_3$	440.422946	439.41567	457.449495
FA 29:0	$C_{29}H_{58}O_2$	438.443681	437.436405	455.47023
FA 29:0;OH	$C_{29}H_{58}O_3$	454.438596	453.43132	471.465145
FA 30:0	$C_{30}H_{60}O_2$	452.459331	451.452055	469.48588
FA 30:0;OH	$C_{30}H_{60}O_3$	468.454246	467.44697	485.480795