

Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies

Appendix 5

Chapter 9 - Compound Identification

Marynka M. Ulaszewska^{1,†}, Christoph H. Weinert^{2,†}, Alessia Trimigno^{3,†}, Reto Portmann^{4,†}, Cristina Andres Lacueva⁵, René Badertscher⁴, Lorraine Brennan⁶, Carl Brunius⁷, Achim Bub⁸, Francesco Capozzi³, Marta Cialiè Rosso⁹, Chiara E. Cordero⁹, Hannelore Daniel¹⁰, Stéphanie Durand¹¹, Bjoern Egert², Paola G. Ferrario⁸, Edith J.M. Feskens¹², Pietro Franceschi¹³, Mar Garcia-Aloy⁵, Franck Giacomoni¹¹, Pieter Giesbertz¹⁴, Raúl González-Domínguez⁵, Kati Hanhineva¹⁵, Lieselot Y. Hemeryck¹⁶, Joachim Kopka¹⁷, Sabine Kulling², Rafael Llorach⁵, Claudine Manach¹⁸, Fulvio Mattivi^{1,19}, Carole Migné¹¹, Linda H. Münger²⁰, Beate Ott^{21,22}, Gianfranco Picone³, Grégory Pimentel²⁰, Estelle Pujos-Guillot¹¹, Samantha Riccadonna¹³, Manuela J. Rist⁸, Caroline Rombouts¹⁶, Josep Rubert¹, Thomas Skurk^{21,22}, Pedapati S. C. Sri Harsha⁶, Lieven Van Meulebroek¹⁶, Lynn Vanhaecke¹⁶, Rosa Vázquez-Fresno²³, David Wishart²³, and Guy Vergères²⁰

¹Department of Food Quality and Nutrition, Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige, Italy

²Department of Safety and Quality of Fruit and Vegetables, Max Rubner-Institut, Karlsruhe, Germany

³Department of Agricultural and Food Science, University of Bologna, Italy

⁴Method Development and Analytics Research Division, Agroscope, Federal Office for Agriculture, Berne, Switzerland

⁵Biomarkers & Nutrimetabolomics Laboratory, Department of Nutrition, Food Sciences and Gastronomy, XaRTA, INSA, Faculty of Pharmacy and Food Sciences, Campus Torribera, University of Barcelona, Barcelona, Spain. CIBER de Fragilidad y Envejecimiento Saludable (CIBERFES), Instituto de Salud Carlos III, Barcelona, Spain

⁶School of Agriculture and Food Science, Institute of Food and Health, University College Dublin, Dublin, Ireland

⁷Department of Biology and Biological Engineering, Food and Nutrition Science, Chalmers University of Technology, Gothenburg, Sweden

⁸Department of Physiology and Biochemistry of Nutrition, Max Rubner-Institut, Karlsruhe, Germany

⁹Dipartimento di Scienza e Tecnologia del Farmaco Università degli Studi di Torino, Turin, Italy

¹⁰Nutritional Physiology, Technische Universität München, Freising, Germany

¹¹Plateforme d'Exploration du Métabolisme, MetaboHUB-Clermont, INRA, UNH, Université Clermont Auvergne, Clermont-Ferrand, France

¹²Division of Human Nutrition, Wageningen University, Wageningen, The Netherlands

¹³Computational Biology Unit, Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige, Italy

¹⁴Molecular Nutrition Unit, Technische Universität München, Freising, Germany

¹⁵Institute of Public Health and Clinical Nutrition, Department of Clinical Nutrition, University of Eastern Finland, Kuopio, Finland

¹⁶Laboratory of Chemical Analysis, Department of Veterinary Public Health and Food Safety, Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium

¹⁷Department of Molecular Physiology, Applied Metabolome Analysis, Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm, Germany

¹⁸INRA, UMR 1019, Human Nutrition Unit, Université Clermont Auvergne, Clermont-Ferrand, France

¹⁹Center Agriculture Food Environment, University of Trento, San Michele all'Adige, Italy

²⁰Food Microbial Systems Research Division, Agroscope, Federal Office for Agriculture, Berne, Switzerland

²¹Else Kröner Fresenius Center for Nutritional Medicine, Technical University of Munich, Munich, Germany

²²ZIEL Institute for Food and Health, Core Facility Human Studies, Technical University of Munich, Freising, Germany

²³Departments of Biological Sciences and Computing Science, University of Alberta, Edmonton, Canada

†First authors

Mol Nutr Food Res DOI: 10.1002/mnfr.201800384

9. Compound Identification

The identification of compounds detected by means of an untargeted analysis is a prerequisite for biological interpretation of the results. While it is comparatively easy to annotate known compounds by library matching in the course of automatic data processing, the identification of yet unknown metabolites is time-consuming and requires a considerable chemical and analytical experience.^[1] In other words, compound identification is still a major bottleneck in metabolomics, especially in view of the fact that a large part, often the majority, of analytes detected by an untargeted analysis is initially unknown. The situation is even worse for food metabolomics, because of the lack of description of many food components in databases and the unavailability of standards.

In order to reduce the identification workload to a reasonable minimum, it is therefore common to initially treat all detected features (either analytes or ions), or at least those that cannot readily be annotated automatically, as non-annotated variables. In the next step, the probably biologically interesting features are then selected by means of statistics or visualization tools. Finally, only the identity of features exhibiting a significant effect is checked, either by verifying the result of the automatic annotation or by performing measures to elucidate their structure. Many tools and approaches are actually developed to improve metabolome annotation and identification of unknown metabolites. However, because of the lack of standardization in the field, multiple bioinformatics solutions with heterogeneous formats are available. This complex landscape not only requires significant resources but also hampers the confidence in the identification process.^[2]

9.1. Levels of Identification

Starting in 2007, the 'Chemical Analysis Working Group' (CAWG) of the Metabolomics Standards Initiative (MSI) defined the best chemical analytical practices in metabolomics. In particular, they proposed minimum reported standards for metadata relative to metabolite identification and defined four levels of metabolite identification:^[3,4]

- Level 1: The compound is identified based on a minimum of two independent and orthogonal data (e.g. mass spectrum and R_t) relative to an authentic compound analyzed under identical experimental conditions. This can be achieved by injecting an authentic standard or by matching the properties of the compound with an in-house spectral library. For most compounds in GC-MS metabolomics, spectral similarity should be at least 90% and the retention index (RI) on the same column type should be within ± 10 -15 RI units or within a relative margin of 0.5-1%^[5]. For the differentiation of closely related isomers with nearly identical spectra (e.g. sugars), a much tighter RI accuracy of about ± 1.0 -1.5 RI units would be needed which is difficult to achieve in practice.
- Level 2: The compound is putatively annotated without chemical reference standards but based on its physicochemical properties and/or its spectral similarity with public/commercial spectral libraries. Spectral matching should be at least 80% and the RI on the same column type should be within ± 15 RI units.

- Level 3: The compound is putatively attributed to a compound class based on its physicochemical properties and/or its spectral similarity with public/commercial spectral libraries.
- Level 4: The compound and its compound class remain unknown. Although unidentified or unclassified, the compound can still be differentiated and (semi-) quantified based on spectral data.

This classification is widely applied in the metabolomic field because of its simplicity. However, in many cases (e.g. the differentiation of isomers), the attribution of a clear identification level is not straightforward.

An alternative option is the quantitative cumulative scoring system, already used in other disciplines of analytical chemistry. This system is based on the use of complementary analytical parameters, such as chromatographic properties (t_R) and spectral properties based on MS (accurate mass, isotopic pattern, in-source fragment, fragmentation spectrum), NMR, IR, or UV. Even if such a scoring can increase the confidence of a structural identification, its application to large data sets is very difficult and resource-intensive so that this approach is actually not applied. In this context, discussions on a revision of the above identification levels are ongoing.^[6,7]

Because of the relatively small number of commercially available metabolite standards that is generally incorporated within an analytical omics methodology, a vast amount of the metabolites that are revealed during an untargeted analysis requires profound efforts to achieve characterization and identification. In some cases, unknown metabolites may actually be known compounds ('known unknowns') with spectra that have often not been made publicly available in accessible libraries. In other cases, the features derived from metabolomics experiments may be truly unknown compounds for which no structure or literature description exists ('unknown unknowns'). The identification of these features often requires a combination of strategies, in which acquisition of both confidence and evidence is necessary.^[8]

9.2. Databases

Biological databanks and databases are important resources for the process of identity annotation. Hereby, a database is defined as "an organized collection of information". It is an aggregation of data and of the computing system used to store them. In comparison, a databank refers to the data only and library is one way of representing information. In bioinformatics, the words databank, database, and library are often used as interchangeable terms.^[9]

Close to 500 databanks and databases are available today,^[10] providing different levels of information and complementary data on chemical structures, physicochemical properties, biological functions, and pathway mapping of metabolites.^[8] The metabolomics community classifies these resources in several categories: i) chemical compound banks; ii) spectral libraries; iii) genome-wide metabolic reconstruction based databases; iv) knowledge databases; and v) references repositories.^[11] Updated inventories of the most useful resources can be found on the portals of the Metabolomics Society (<http://metabolomicsociety.org/resources/metabolomics-databases>) and the Food Biomarker Alliance (<http://foodmetabolome.org/wpkg4>). The choice of database(s) during the process of identification is not

straightforward, although acknowledged to be significantly impacting the quality of the annotation results and, consequently, the quality of the biological discussion that will arise from these findings. The next section provides an overview of a selection of the databanks and databases mostly used for nutritional metabolomics.

9.2.1. Chemical Compounds Centric Resources

- The *Chemical Entities of Biological Interest* (ChEBI) (<http://www.ebi.ac.uk/chebi/>)^[12] is a public and reference dictionary of molecular entities focused on 'small chemical compounds' hosted by the European Bioinformatics Institute (EMBL-EBI). The 41,000 entities are manually annotated by the ChEBI team providing high-quality data to researchers. The ChEBI ontologies provide molecular structures, biological functions, and subatomic particle information. A manual search is available by mass and chemical formula. Web services for programmatic access are also provided.
- *KNApSack* (<http://kanaya.naist.jp/KNApSack/>), supported by the Kanaya laboratory and the Nara Institute of Science and Technology (NAIST) in Japan, is an extensive plant species-metabolite database, which includes over 20,000 edible and non-edible plants and contains over 50,000 metabolites. The main purpose of this database is the accumulation and search of metabolite-species relationships. The database provides manual and automatic searching utilities supporting MS peak, molecular weight, chemical formula, and species filters.

9.2.2. Spectral Centric Resources

- *MassBank*^[13] is a database of comprehensive, high-resolution mass spectra of metabolites supported by the Japan Institute for Bioinformatics Research and Development. The web portal provides spectral and structural searching utilities as well as web services for free access (<http://www.massbank.jp/>). This public and reference resource contains more than 41,000 MS spectra for 47,000 compounds. The project is based on distributed databases by a large network of contributors providing annotated spectra and their metadata. Of note, significant work on controlled vocabulary and format was conducted to collect harmonized information (http://www.massbank.jp/manuals/MassBankRecord_en.pdf).
- *METLIN* (<http://metlin.scripps.edu/>)^[14] is a metabolite database focusing (although not exclusively) on metabolomic approaches. This database contains information about more than 64,000 structures (pharmaceuticals and their metabolites, pesticides, human and animal endogenous compounds, peptides, plant metabolites, etc.). METLIN incorporates more than 59,000 high-resolution MS/MS spectra at different collision energies.
- *mzCloud* (<https://www.mzcloud.org/>) is a database of high-resolution tandem mass spectra, which are arranged into spectral trees obtained from Orbitrap mass spectrometers. MS/MS and multi-stage MSⁿ spectra were acquired at various collision energies, precursor *m/z*, and isolation widths using CID as well as higher-energy collisional dissociation (HCD). mzCloud is a fully searchable library, which allows a range of searches on

spectra, trees, structures and substructures, monoisotopic masses, peaks (*m/z*), precursors, and molecular names.

- The *Golm Metabolome Database* (GDM) (<http://gmd.mpimp-golm.mpg.de/>)^[15] is an open access metabolome database dedicated to GC-MS users, which provides public access to custom mass spectral libraries, metabolite profiling experiments as well as additional information and tools, e.g. with regard to methods, spectral information, or compounds.
- *The Fiehn library* (<http://fiehnlab.ucdavis.edu/19-projects/153-metabolite-library/>)^[16] is a collection of mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight GC/MS covering lipids, amino acids, fatty acids, amines, alcohols, sugars, amino-sugars, sugar alcohols, sugar acids, organic phosphates, hydroxyl acids, aromatics, purines, and sterols. The libraries can be used in conjunction with GC/MS software but also support compound identification in the public of metabolites from plants, animals, and microorganisms.
- The *NIST* library is a commercial spectral database of the US National Institute of Science and Technology (<https://www.nist.gov/srd>). Its last version NIST14 contains 234,284 ESI MS/MS spectra of small molecules, including authentic chemical standards of metabolites, lipids, biologically active peptides, and all di-peptides and tryptic tripeptides, as well as a collection of 276,259 EI mass spectra from 242,477 unique compounds.

The characteristics of mass spectral libraries have been reviewed recently^[17] and, due to their complementary nature, their combined use is recommended. However, despite their considerable size, all of the mentioned spectral libraries are far from complete and they are growing, if at all, rather slowly. Further, although EI spectra and RI are generally highly comparable in GC analysis and Orbitrap spectra are highly reproducible and easily comparable with those in mzCloud, subtle but sometimes relevant differences between different types of instruments or even between chemically very similar stationary phases from different column manufacturers, respectively, can be observed for LC-MS/MS. For this reason, the development of in-house spectral libraries is always the most reliable (but also most expensive) way to ensure highly reliable automatic annotation as well as support for manual compound identification.

9.2.3. Knowledge Databases

- *The Human Metabolome Database* (HMDB)^[18] supported by the Canadian Metabolomics Innovation Centre, contains detailed information about more than 110,000 small molecules/metabolites found in the human body. This public and reference resource uses a metabocard system to provide the following information: i) chemical and biochemistry properties; ii) clinical data; iii) several "cross web links" with other reference resources; and iv) downloadable MS and NMR spectra acquired from various instruments in low and high resolution. When experimental spectra are not available, HMDB 4.0 proposes *in silico* predicted spectra. A manual search by name, spectra or structure is available.
- *FoodDB* (www.FoodDB.ca) is the most comprehensive resource on the chemistry and biology of food constituents. It provides information on both

macronutrients and micronutrients, including many of the constituents that give foods their flavor, color, taste, texture, and aroma. Each chemical entry in FooDB contains more than 100 separate data fields covering detailed physico-chemical, biochemical, and spectral information as well as food composition data.

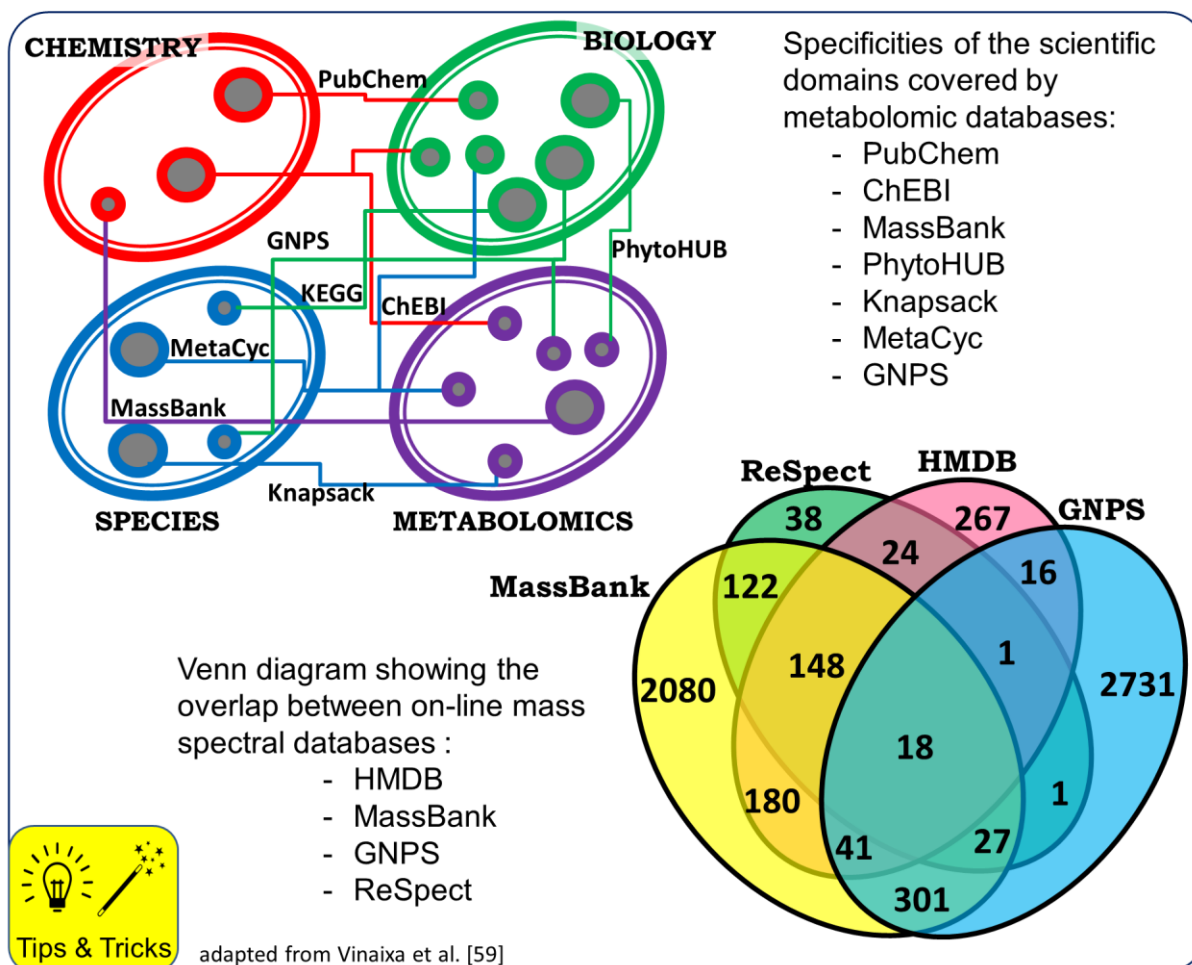
- *PhytoHUB* (www.phytoHUB.eu) inventories all dietary phytochemicals present in commonly consumed foods and their human metabolites. It provides data on the main dietary sources, selected literature references on human metabolism, MS spectra and has search possibilities designed for use in nutritional metabolomics studies. PhytoHub is manually curated by an international team of scientists with expertise on phytochemicals.

9.2.4. Reference Repositories

With more than 300 available studies, 25,000 metabolites identified in 2,800 species, MetaboLights (MTBLS) is a public reference repository for metabolomics experiments and derived information. Described as a cross-species, cross-techniques resource,^[19] MTBLS covers metabolite properties and their experimental fingerprints as well as their potential functions in pathways and biological systems.^[20] The web portal provides full searching utilities to explore studies, metabolites, species data as well as spectral similarities and chemical structure tools. Based on community contributions, the repository offers submission tools and methodologies based on a high quality controlled vocabulary and semantics.

Table S9.1 gives an overview of the most relevant databases available in the field of metabolomics of which some have been discussed above in more detail in order to point out the different database structures. When choosing the databases for the annotation step, scientists need to consider both the size and the specificities of the information contained in these databases (**Box S9.1** below). Databases extracted for model organisms, which usually contain smaller datasets, can be used in priority to find more relevant identity hypotheses, for example HMDB for human studies, PhytoHub for human studies with consumption of plant foods, KNApSACk for plant-related studies. Large databases such as PubChem^[21] (157,000,000 entries) will always propose a higher number of possible molecular identifications, but they will include many false positives.

As no single database can address all issues related to annotation, the usage of a number of them can be a good strategy to increase annotation efficiency. When using several databases to drive the annotation step, scientists should take into consideration the possible overlap between them. Indeed, most repositories use mirroring processes, i.e. duplicate existing data to their own data repository, artificially increasing confidence by the simple fact that hypothetical metabolites are presented by several databases. This point was first highlighted in 2016 by Vinaixa et al.,^[8] who calculated the overlap between open mass spectral databases such as HMDB, MassBank, and ReSpect (see **Box S9.1**).



Box S9.1. Set of databases and on-line libraries used in metaoblimomics.

Table S9.1. Overview of metabolomic databases.

| | |
|---|---|
| BioMagResBank (BMRB) | http://www.bmrwisc.edu/metabolomics/ |
| Birmingham Metabolite Library Nuclear Magnetic Resonance Database | http://www.bml-nmr.org/ |
| Cerebrospinal Fluid (CSF) metabolome database | http://www.csfmetabolome.ca/ |
| Chemical Entities of Biological Interest (ChEBI) | www.ebi.ac.uk/chebi/ |
| Chemspider | http://www.chemspider.com/ |
| DrugBank | http://www.drugbank.ca/ |
| FooDB | http://foodb.ca/ |
| GNPS Spectral search | http://gnps.ucsd.edu |
| Golm Metabolite Database (GMD) | http://csbdb.mpimp-golm.mpg.de/csbdb/home/databases.html |
| Human Metabolome Database (HMDB) | http://www.hmdb.ca/ |
| HumanCyc | http://humancyc.org/ |
| KNAPSAck | http://kanaya.naist.jp/KNAPSAck/ |
| Kyoto Encyclopedia of Genes and Genomes (KEGG) | http://www.genome.jp/kegg/ |
| Lipid Maps Structure Database (LMSD) | http://www.lipidmaps.org/ |
| LipidHome | https://www.ebi.ac.uk/metabolights/lipidhome/ |
| Madison Metabolomics Consortium Database | http://mmcd.nmrwisc.edu/ |
| MassBank | http://www.massbank.jp/ |
| Matador | http://matador.embl.de |
| MetaboLights (MTBLS) | https://www.ebi.ac.uk/metabolights/ |
| MetaCyc | http://metacyc.org/ |
| METLIN | http://metlin.scripps.edu/index.php |
| MMCD Spectral search | http://mmcd.nmrwisc.edu/ |
| MoNA Spectral search | http://mona.fiehnlab.ucdavis.edu/ |
| MyCompound ID | www.mycompoundid.org |
| mzCloud | https://www.mzcloud.org |
| National Institute of Science and Technology (NIST) | http://www.sisweb.com/software/ms/nist.htm |
| Phenol-explorer | phenol-explorer.eu |
| PhytoHUB | http://phytohub.eu/ |
| Plant Metabolome Database (PMDB) | http://scbt.sastra.edu/pmdb/ |
| PredRet | predret.org |
| PubChem | http://pubchem.ncbi.nlm.nih.gov/ |
| Reactome | http://www.reactome.org/ |
| ReSpect | http://spectra.psc.riken.jp/ |
| Retention prediction | www.retentionprediction.org |
| Search Tool for Interaction of Chemical (STITCH) | www.retentionprediction.org |
| SuperTarget | http://stitch.embl.de/ |
| Therapeutic Target Database (TTD) | http://bioinf-apache.charite.de/supertarget_v2/ |

9.3. Identification of Compounds in GC/MS-based Metabolome Studies

9.3.1. Fragmentation Patterns

The fragmentation patterns of TMS or MTS derivatives of the compound classes that are commonly detected in a GC-based metabolomics analysis have been studied since the late 1950s.^[22] Although especially the mass spectrometric characteristics of sugars,^[23-29] sugar alcohols and inositols,^[30-32] amino acids and amines,^[33-36] aliphatic and

aromatic acids,^[37-42] as well as purines and pyrimidines^[43] have been studied in some detail, there is still much room for further investigations in order to obtain a better understanding of the fragmentation mechanisms and to find marker fragments for specific compound classes, functional groups, or positional isomers. This section aims to give an overview of the existing knowledge.


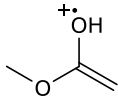
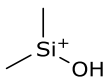

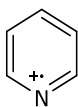
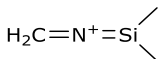
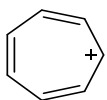
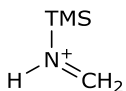
The intensity of fragmentation of TMS or MTS derivatives varies greatly. Most of the aliphatic derivatives show fragmentation to such an extent that the molecular ion [M] is of low intensity or cannot be observed at all. However, the

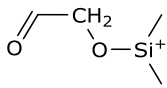
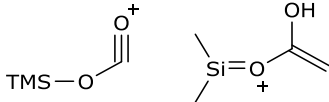
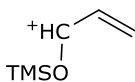
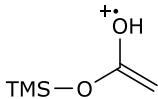
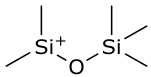
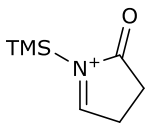
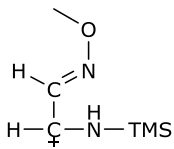
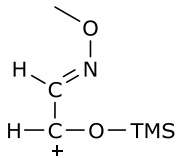
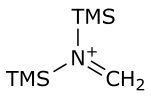
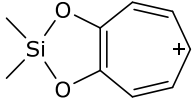
loss of a methyl radical from the molecular ion [M-15] can usually be observed and some helps to identify the molecular ion.^[44, 45] In contrast, the spectra of aromatic and especially phenolic compounds exhibit an intense molecular ion and less fragmentation^[40] due to the possibility to stabilize the molecular ion by charge delocalization.

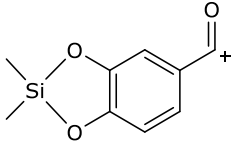
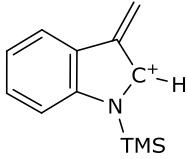
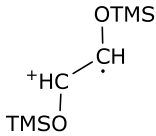
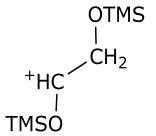
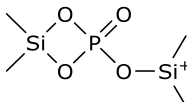
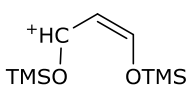
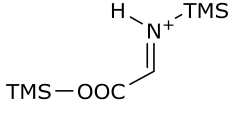
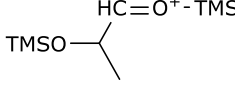
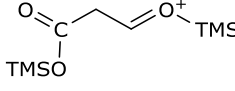
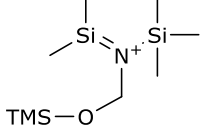
The spectra of silylated compounds nearly always exhibit a prominent fragment with m/z 73 and, typically if at least two TMS groups are present in the molecule, also the fragment m/z 147. While these fragments are not very useful for detailed structure elucidation, their occurrence proves at least the presence of silylated functional groups while their absence indicates that no silylable groups exist. Beyond that, often many other ions are formed after EI of derivatized metabolites. While their structures could be elucidated in

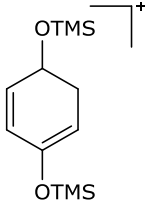
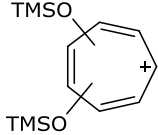
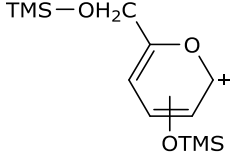
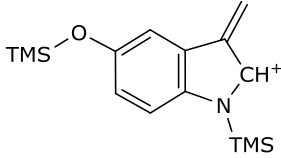
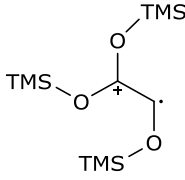
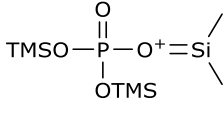
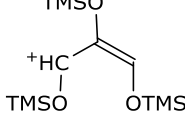
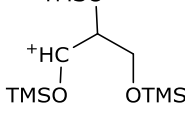
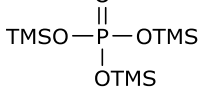
many cases, it is generally true that only a combination of fragments rather than the occurrence of single marker fragments enables unequivocal compound identification. This is the reason why the availability of comprehensive spectral libraries is highly important. Nevertheless, a detailed knowledge about the fragmentation mechanisms behind the EI spectra of metabolite derivatives is crucial for the identification of unknowns - as the existing spectral libraries are still incomplete, compared to the vast number of naturally occurring compounds. In the following, a brief overview of the mass spectrometric features of major classes of derivatized metabolites will be provided and the proposed structures of a range of common fragments are provided in **Table S9.2**.

Table S9.2. Possible structures of fragment ions frequently occurring in the spectra of TMS or MTS derivatives of biological metabolites.

| m/z | Proposed Structure | Remarks / Occurrence | Literature |
|-----------|---|---|--------------|
| 73.04680 |  | Unspecific fragment | [24, 34, 45] |
| 74.03623 |  | Fatty acid methyl esters | [46, 47] |
| 75.02607 |  | Unspecific fragment | [24, 34, 45] |
| 77.03858 |  | Unspecific fragment | [48] |
| 79.04165 |  | Pyridine solvent (often as tailing) | [37] |
| 86.04205 | $\text{H}_2\text{C}=\text{N}^+=\text{Si}$  | Amines; derived from 174.11288 | [33, 35] |
| 91.05423 |  | Phenols | [46] |
| 100.05770 | $\text{HC}\equiv\text{N}^+ - \text{TMS}$ | Amines, derived from m/z 102.07335 or 174.11288 | [33, 36] |
| 102.07335 |  | Primary amines | [33, 36] |
| 103.05737 | $\text{TMS}-\text{O}^+=\text{CH}_2$ | Unspecific sugar fragment | [23, 25, 31] |

| | | | |
|-----------|---|---|------------------|
| 117.07302 | $\text{H}_3\text{C}-\underset{\text{H}}{\text{C}}=\text{O}^+-\text{TMS}$ | Deoxy sugars or polyols | [27, 32] |
| 117.03663 |  | Non-methoximated TMS derivatives of aldoses (C5-C6) | [23] |
| 117.03663 |  | TMS esters of carboxylic acids | [33, 34] |
| 129.07302 |  | Unspecific sugar fragment | [23] |
| 132.06011 |  | Fatty acid TMS esters | [42] |
| 147.06559 |  | Unspecific fragment of derivatives with at least 2 TMS groups | [24, 33, 34, 45] |
| 156.08392 |  | Several amino acids | [34] |
| 159.09482 |  | Amino sugars | |
| 160.07883 |  | C2 fragment of aldoses; absent in 2-ketohexoses | [25, 27] |
| 174.11288 |  | Aliphatic primary amine derivatives; also some amino acids | [34, 36, 45] |
| 179.05228 |  | Phenolic acids with o-dihydroxy structure | [36, 40] |
| 191.09181 | $\text{TMS}-\text{O}^+=\underset{\text{H}}{\text{C}}-\text{OTMS}$ | Unspecific sugar fragment | [23, 24, 31] |

| | | | |
|-----------|---|---|----------------------|
| 193.03155 |  | <i>o</i> -Dihydroxybenzoic acids | [40] |
| 202.10465 |  | Indoles; tryptophan | |
| 204.09963 |  | Sugars; predictive for pyranoside structure | [23, 24, 28-31, 41] |
| 205.10746 |  | Sugars | [23, 25, 27] |
| 211.00062 |  | Phosphorylated compounds | [24, 26] |
| 217.10746 |  | Sugars; form by TMSOH loss from <i>m/z</i> 307 | [23, 24, 30, 31, 41] |
| 218.10271 |  | Amino acids | [33, 34, 45] |
| 219.12311 |  | Deoxy sugars and polyols | [32] |
| 233.10237 |  | 2-Deoxy acids like malic acid and related C5/C6 acids | [38] |
| 248.13167 |  | Glycine; β -alanine and GABA | [34] |

| | | | |
|-----------|---|---|--------------|
| 255.12311 |  | Quinic acid and derivatives | [41] |
| 267.12311 |  | Some phenolic acids with dihydroxybenzene structure | [36, 37] |
| 271.11802 |  | Disaccharides and glycosides; formed by loss of TMSOH from <i>m/z</i> 361 | [23] |
| 290.13909 |  | Hydroxyindoles | [34] |
| 292.13408 |  | Organic acids with 2,3-dihydroxy structure; formed by rearrangement | [38] |
| 299.07145 |  | Phosphorylated compounds | [24] |
| 305.14190 |  | Prominent with inositols | [23, 24, 31] |
| 307.15755 |  | Sugars; especially C6 ketoses and C5 sugars | [27, 30] |
| 314.09493 |  | Molecular ion of trimethylsilyl phosphate | [24] |

| | | | |
|-----------|--|---|----------|
| 315.10275 | $\begin{array}{c} \text{OH} \\ \\ \text{TMSO}-\text{P}=\text{O}^+-\text{TMS} \\ \\ \text{OTMS} \end{array}$ | Phosphorylated compounds | [24] |
| 318.14973 | $\begin{array}{c} \text{TMS}-\text{O}^+=\text{CH} \\ \\ \text{C}=\text{C} \\ \quad \\ \text{OTMS} \quad \text{H}-\text{OTMS} \end{array}$ | Inositols | [31] |
| 319.15755 | $\begin{array}{c} \text{TMSO} \quad \text{OTMS} \\ \quad \\ \text{HC} \quad \text{C} \\ \quad / \quad \backslash \\ \text{OTMS} \quad \text{C} \end{array}$ | Methoxime-TMS and TMS derivatives of sugars; derived from m/z 409 by elimination of TMSOH | [23, 27] |
| 321.17320 | $\begin{array}{c} \text{TMS}-\text{O}^+=\text{CH} \\ \\ \text{C} \\ \quad \\ \text{TMSO} \quad \text{OTMS} \end{array}$ | Deoxy sugars | [32] |
| 333.13682 | $\begin{array}{c} \text{OTMS} \\ \\ \text{C} \\ / \quad \backslash \\ \text{TMSO} \quad \text{O} \\ \quad \\ \text{OTMS} \end{array} \quad \begin{array}{c} \text{TMSO} \quad \text{CHO} \\ \quad \\ \text{C} \quad \text{C} \\ / \quad \backslash \\ \text{OTMS} \quad \text{OTMS} \end{array}$ | C6 sugar acids (C3-C6); derived from m/z 423 by elimination of TMSOH; often combined with m/z 292 | [38] |
| 345.17320 | $\begin{array}{c} \text{OTMS} \\ \\ \text{C} \\ \quad \\ \text{TMSO} \quad \text{OTMS} \end{array} \quad \text{---}^+$ | Quinic acid and derivatives | [41] |
| 357.11332 | $\begin{array}{c} \text{OTMS} \\ \\ \text{O} \\ / \quad \backslash \\ \text{P}^+ \\ \quad \\ \text{TMSO} \quad \text{O} \end{array}$ | Phosphorylated compounds | [26] |
| 361.16812 | $\begin{array}{c} \text{TMS}-\text{OH}_2\text{C} \\ \\ \text{C} \\ \quad \\ \text{TMS}-\text{O} \quad \text{O} \\ \quad \\ \text{TMS}-\text{O} \end{array} \quad \text{---}^+$ | Disaccharides and glycosides; derived from m/z 451 by elimination of TMSOH | [23, 29] |
| 387.14228 | $\begin{array}{c} \text{---TMS} \\ \\ \text{O}^+ \\ \\ \text{TMSO}-\text{P}-\text{OTMS} \\ \\ \text{OTMS} \end{array}$ | Phosphorylated sugars | [24] |

| | | | |
|-----------|--|---|------|
| 437.20256 | | Non-methoximated TMS derivatives of aldoses; some disaccharides | [23] |
| 451.21821 | | Glycosides and disaccharides; usually in low intensity | [29] |

- The spectra of sugars, sugar alcohols, amino sugars and inositols contain, at first glance, not much structural information. The molecular ion, the [M-15] ion, and other large fragments are often extremely weak or too labile to be detected under EI conditions. Instead, there is a comparatively small number of ions that are prominent in many spectra, like m/z 103, 117, 160, 204, 205, 217, 307, 319 and 361. These fragments are predominantly formed by α -cleavage of the carbon chain and elimination of TMS by rearrangement.^[30] The patterns of these main fragments in combination with the retention indices enable at least a distinction between the above mentioned main classes. Beyond this, some fragments allow drawing a direct conclusion about the structure of the derivative. For example, m/z 160 is a typical C2 fragment of 2-aldoses like glucose or galactose but virtually absent in 2-ketoses like fructose; in case of non-acetylated amino sugars like glucosamine, m/z 159 instead of m/z 160 is present. In case of non-oximated TMS derivatives of sugars, m/z 204 is a predictor for a pyranoside structure^[24] but can also be found in the spectra of other compounds like quinic acid 5TMS and serine 3TMS. In some cases, deoxy sugars like fucose or rhamnose or deoxy sugar alcohols can be recognized with the help of the fragment m/z 117.07302 (although this requires high-resolution measurement due to isobaric ions; see below), with m/z 219 and 321 being further specific but rather minor fragments. m/z 361 indicates a glycoside structure and is prominent in the spectra of disaccharides. The fragments m/z 305 and 318 are specific for inositols.^[31] However, a distinction between the different stereoisomers of one of the above-mentioned compound classes solely based on the spectra is virtually impossible and chromatographic resolution is crucial.
- Aliphatic organic acids share many fragments with the sugars and sugar alcohols but their spectral footprints are in general more unique due to their higher structural diversity. For this reason, spectral matching is often successful but even unknown organic acids can be recognized as members of this compound class, based on general spectral similarity but also some marker fragments. For example, m/z 292 indicates the existence of two hydroxyl groups in α - and β -position to the carboxyl group. Further, the spectra of the so-called sugar acids like gluconic acid, glucaric acid, glucuronic acid and glucoheptonic acid as well as their isomers are more homogenous but exhibit the specific fragment m/z 333.^[38] Moreover, the alicyclic compound quinic acid and related conjugates (e.g. chlorogenic acids) can unambiguously be identified using the fragments m/z 255 and 345.
- Aromatic acids like phenolic acids form comparatively stable [M] and [M-15] ions. Usually, neutral losses characterize the further fate of these ions but also fragments indicating specific substructures like m/z 179, 193, 267, and others can sometimes be observed.^[36, 37, 40] As even closely related isomers of aromatic acids can usually be separated by chromatography and exhibit also distinct spectra, identification of aromatic acid derivatives is often comparatively straightforward. The situation is similar for TMS derivatives of purines and pyrimidines.^[43]
- The molecular structures of the amino acids are diverse and thus the spectra of their TMS derivatives are heterogeneous. With a few exceptions, the fragment m/z 218 which represents the 'head group' of N-(trimethylsilyl)trimethylsilyl ester derivatives after loss of the side chain^[33, 34, 45] is commonly observed but with varying intensity. Other fragments are more specific, e.g. m/z 156 (a characteristic cyclization product formed from the derivatives of glutamic acid, acetylglutamic acid, glutamine, and oxoproline), m/z 248 for glycine, β -alanine and GABA or m/z 202 and 290, which indicate indole or hydroxyindole substructures, respectively. Because the mechanisms that govern the fragmentation of amino acid derivatives are diverse and have not been investigated comprehensively so far, the structure elucidation of unknown amino acids can be time-consuming.
- The spectra of the TMS derivatives of aliphatic amines are often not very informative. The unspecific fragments m/z 174 or 102 (representing di- or mono-trimethylsilylated primary amino groups, respectively) as well as their breakdown products usually dominate the spectra and indicate the existence of amino groups, but other informative fragments are often very weak or missing due to the generally intense fragmentation.^[33, 34, 36, 45]
- Unique fragments are observed in the spectra of phosphorylated compounds, in addition to the fragments corresponding to the non-phosphorylated metabolite backbone. Besides the common ions m/z 299 and 315, the fragments m/z 211, 357 and 387 may occur.^[24, 26]
- The members of other compound classes that may be detected in an untargeted metabolome analysis (independent of the question whether they are process artefacts or indeed originating from the biological material of interest) can often be identified based on specific ions, for example fatty acid methyl esters: m/z 74 and 87; fatty

acid trimethylsilyl esters: m/z 117 and 132; n-alkanes: m/z 57, 71 and 85; cyclosiloxanes: m/z 207, 281 and 355 (among others); phthalates: m/z 149.

It has to be mentioned that EI spectra of TMS derivatives of biological compounds have to be interpreted with caution for several reasons. Firstly, as the example of the ion m/z 117 in **Table S9.2** shows, isobaric ions which cannot be distinguished using low-resolution instruments may occur, so the specificity of marker ions cannot be validated solely based on nominal mass data. However, if different ions with the same sum formula (as in case of m/z 117.03663) may occur, even high-resolution measurements may be insufficient. This means that the occurrence of specific marker ions always has to be interpreted in view of the other fragments in the spectrum. Further, it has been shown that generally all kinds of rearrangements may occur during the fragmentation of TMS derivatives, including the intra- and intermolecular transfer not only of protons but also of TMS groups,^[23, 45, 49] which may further complicate the interpretation of the spectra. Finally, side products and artefacts may already be formed during derivatization,^[50] with the formation of non-methoximated TMS derivatives of sugars being an example especially relevant for high-sugar matrices.^[51]

In summary, although EI spectra are highly comparable and thus ideal for library matching, the generally intense fragmentation of the metabolite derivatives may hamper spectral interpretation and compound identification. Therefore, further tools should be taken into consideration. The use of deuterated silylation reagents is one potentially effective approach. Further, CI (see section 6.1.1) or APCI^[52-54] may help to reduce in-source fragmentation, to identify the molecular ion and to calculate elemental composition.

Table S9.2 gives an overview of possible structures of fragments that occur commonly in the spectra of TMS or MTS derivatives of biological metabolites, which have been proposed in literature. If available, also alternative structures are provided. If no structures were found in the literature, own proposals were made based on the author's own spectral collection.

9.3.2. Retention Indices and Retention Time

As mentioned in sections 5.3, 5.4 and 6.1.1., the EI spectra of different compounds can be similar or nearly identical, especially in case of stereoisomers. Although sometimes difficult, a chromatographic separation of such compounds is often achievable. Nowadays, GC capillary columns can be manufactured in a highly standardized way and the separation of metabolite derivatives on the apolar stationary phases commonly used in GC-based metabolomics is, with a few exceptions, highly reproducible, even if columns from different manufacturers are compared. This opens the possibility of using measures of relative retention as a second identification criterion in addition to the mass spectrum. The RI concept was invented in 1958 by Kováts for isothermal conditions^[55] and later modified for linear temperature-programmed GC by van den Dool and Kratz.^[56] Homologous series of organic compounds, typically n-alkanes, are used as reference. The indices of the RI marker compounds are arbitrarily defined by multiplying the carbon number by hundred. After analyzing the target compounds under the same conditions, the RIs are calculated by local interpolation using a simple equation^[56] based on the t_R of the analyte as well as the t_R and the pre-defined RI of the RI markers.

In principle, any homologous series of GC-compatible organic compounds can be used as RI markers.^[57, 58] In the metabolomics community, however, only n-alkanes and FAMES are used. While the RI data for n-alkanes are clearly more comprehensive, the n-alkanes have the practical drawback that the spectra (especially of the long-chain alkanes) are virtually identical and can thus be easily mixed up, not only in case of automatic data processing. In the spectra of FAMES, the molecular ion can always be observed which facilitates annotation. Although the use of FAMES as RI markers is thus advantageous, it has not yet gained wider acceptance due to the lacking of RI reference data.^[59] A further limitation is that very long FAMES with a chain length of $> C30$ are, at the time of writing, commercially unavailable. Among the currently available spectral libraries used in the metabolomics community, only the Fiehn library uses FAME as RI markers by default.^[16] In any case, prior to routine application, the investigated samples must carefully be checked for endogenous occurrence of chosen RI compound series.

In theory, the formula suggested by van den Dool and Kratz^[56] is only valid if the GC temperature program is strictly linear. This means that any deviation from linearity, i.e. different slopes or even plateaus, can lead to errors.^[57] Practical experience shows that this is especially true for the early-eluting metabolites. Further, the presence of a biological matrix may deteriorate the accuracy of RI determination and RI calculation by polynomial regression may be more precise than the traditional van den Dool approach.^[5]

Attempts have also been made to adapt the concept of RI for both chromatographic dimensions in GC \times GC.^[60-62] On the one hand, RI determined on a single column in one-dimensional GC are nearly consistent with those obtained for the 1D of a GC \times GC column combination, given that the stationary phases are comparable. On the other hand, the practical implementation of an RI calibration for the 2D would be complex and tedious and it is thus questionable if the afore-mentioned concepts will gain wider acceptance in the metabolomics community as well as other application fields for GC \times GC.

In conclusion, RI are an important identification criterion and their use can be considered as mandatory in GC-based metabolomics, especially for isomer annotations. The comparability of RI is generally high but deviations may occur due to slight difference between stationary phases as well as different temperature programs.

9.3.3. Spiking With Standards

The gold standard concerning compound identification is of course the injection of a reference compound. As chemical suppliers expand their product portfolio continually, more and more analytical standards become commercially available. Further, efforts have been made to intensify the academic exchange of rare analytical standards via for example the collaborative exchange platform FoodComEx (Food Compounds Exchange, www.foodcomex.org) developed within the scope of the JPI project FoodBAIL.^[63]

As the biological matrix may have an impact on the retention of an analyte in a GC system, it is advisable to perform the final identification by comparing the runs of a matrix sample with and without spiking of the pure compound and, if possible, some closely related compounds to avoid errors due to a hidden lack of selectivity. For example, the separation of stereoisomeric sugars, which

typically exhibit (nearly) identical EI spectra is often incomplete in one-dimensional (see sections 5.3 and 9.3.2) or even two-dimensional GC (see section 5.4). Thus, even co-chromatography with an authentic standard can lead to a false annotation, especially in case of stereo isomers of for example all kinds of sugars and sugar alcohols in complex chromatogram regions. A traditional but recommendable approach is to verify compound identity using another stationary phase.

9.3.4. Computational Approaches

The identification of high numbers of metabolites from GC-MS profiling experiments is actually a major challenge, especially for novel compounds (the 'unknown unknowns'), which requires the use of computer tools for structure elucidation.^[64] "Computational mass spectrometry" deals with the development of computational methods for the automated analysis of MS data. The first aspect of information, which can be extracted from accurate mass, is the elemental composition. Different software programs based on isotope pattern analysis are available, either to simulate them, decompose monoisotopic peaks, i.e. find all molecular formulas sufficiently close to the measured mass, or score candidate compounds.^[65] Moreover, to exclude possible molecular formulas, different approaches were proposed, like the 'seven golden rules' by Kind et al.^[66] based on a chemistry basis.

However, if a compound is unknown, comparing its fragmentation spectrum to a spectral library or interpreting it manually remains a major challenge: it requires expert knowledge and is very time consuming. Therefore, particularly in metabolomics, *in silico* fragmentation solutions are used. Different computational approaches are actually available:

- **Mass spectral classifiers:** as EI fragmentation is well understood, many mass spectral classifiers have been provided in the field. A feature-based classification approach was proposed by Varmuza et al.^[67] for EI spectra. It uses a set of classifiers to identify the presence of 70 substructures in a compound. More recently, Hummel et al.^[68] subdivided the Golm Metabolome Database into several classes and proposed a decision tree-based prediction of the most frequent substructures, based on mass spectral features and RI information to classify unknown metabolites into different compound classes.
- **Rule-based spectrum prediction:** these algorithms are based on rules extracted from the MS literature over years. Different commercial solutions are actually available. Initially designed for the prediction and interpretation of fragmentation after EI, there are also used to interpret CID fragmentation. Mass Frontier (HighChem, Ltd. Bratislava, Slovakia; versions after 5.0 available from Thermo Scientific, Waltham, USA), including a large manually curated fragmentation library, allows the prediction of spectra from hypothesized structures or compounds of interest. CD/MS Fragmenter (Advanced Chemistry Labs, Toronto, Canada) proposes to interpret a given fragmentation spectrum using a known molecular structure. Molgen-MS^[69] uses the same approach but also accepts additional fragmentation mechanisms.

- **Combinatorial fragmentation:** these algorithms use bond disconnection to explain peaks in the fragmentation spectrum. It is important to note that this approach is not proposing fragments resulting from rearrangements. Recently, MetFrag^[70] was developed to allow the screening of thousands of candidates from compound databases and to rank them depending on the agreement between the experimental and the *in silico* fragments. It was extended to analyze EI fragmentation and include other analytical criteria (like retention behavior).^[71]
- **Fragmentation trees:** this approach consists in computing a fragmentation tree that explains the peaks in the fragmentation spectrum. These algorithms neither need spectral libraries nor molecular structure databases and are therefore an interesting way of identifying 'unknown unknowns'. Fragmentation trees were first applied for metabolite identification from multistage MS data,^[72] but it was also used for *de novo* analysis of EI mass spectra.^[73]

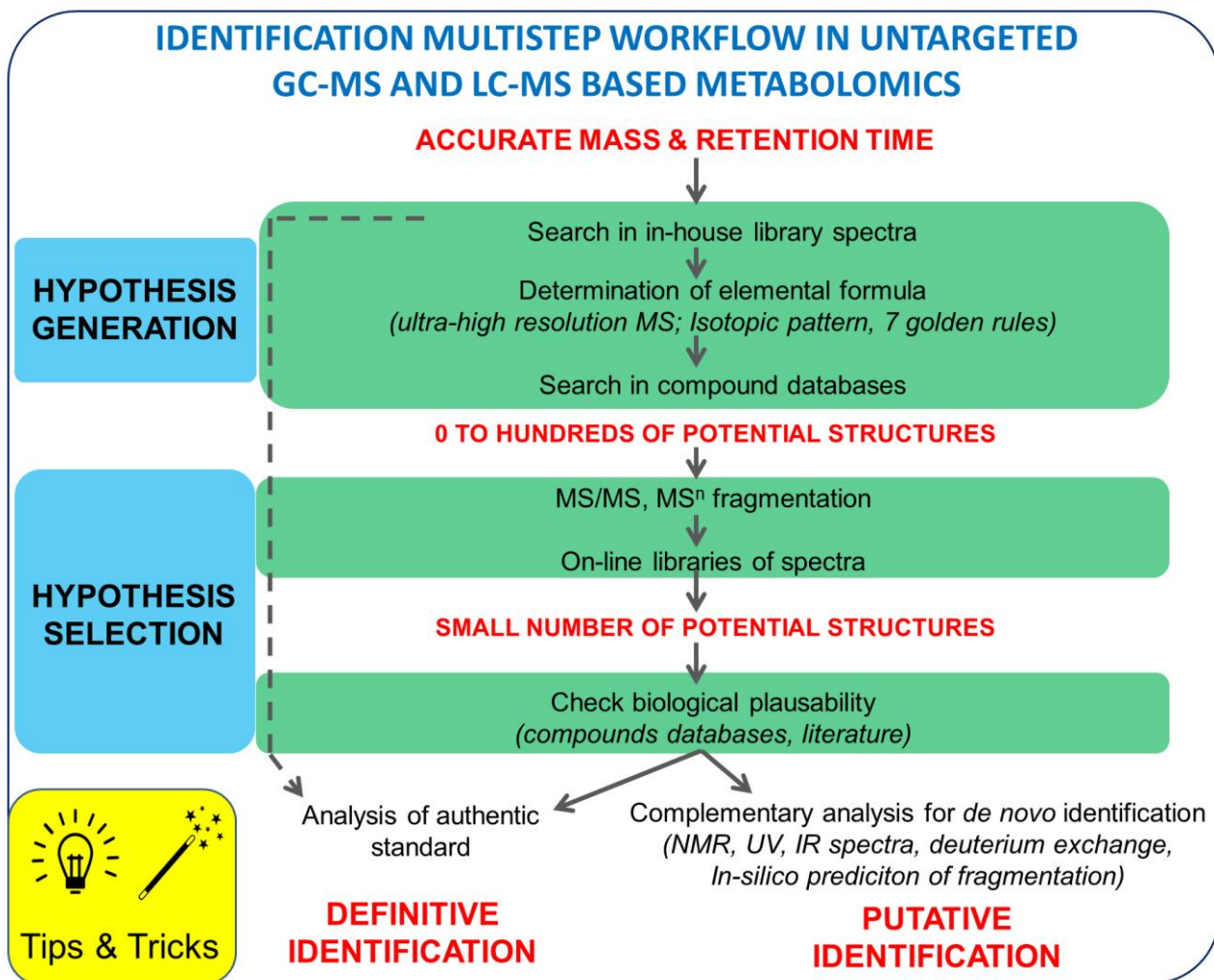
9.4. Identification of Compounds in LC/MS-based Metabolome Studies

Untargeted metabolomics comprises the comparison of the relative abundances of metabolites in multiple samples without prior identification. After the data processing step, including statistical analyses, a list of interested features, characterized by their R_t and mass spectral data (accurate mass, isotopic pattern, adducts, in-source fragments) are candidates for identification. In this chapter, we will describe the full strategy starting from acquiring information for these peaks until the structural identification of potential biomarkers.

The first step of the workflow will focus on generating hypotheses, i.e. determining the elemental formulae and querying in compound databases. Then the second step will consist in accumulating proofs of structural evidence regarding potential structures. In the last step, either authentic standards will allow the validation of the structural identification, or other complementary analyses will be performed to bring new knowledge about unknown metabolites. A general workflow for the annotation step is given in **Box S9.2**, while a real example of annotation for LC-MS based metabolomics is shown in **Boxes S9.3 and S9.4**.

9.4.1. Hypothesis Generation

The first step in the annotation process is the calculation of the elemental composition of a feature of interest, considering information about the isotopic pattern, adducts, and fragments present in the full scan spectra. Protonated and deprotonated molecule species are predominant in the mass spectra, however, ions with adducts such as sodium, potassium or more complex additive species of formiate, acetate or ammonium are commonly found too, as they originate from mobile phases, buffers, or samples in se. The in-source fragmentation is a natural phenomenon in LC-MS and, unlike adducts and neutral losses, the exact mass of in-source fragments is sometimes difficult to predict. Verification of in-source fragments must be performed

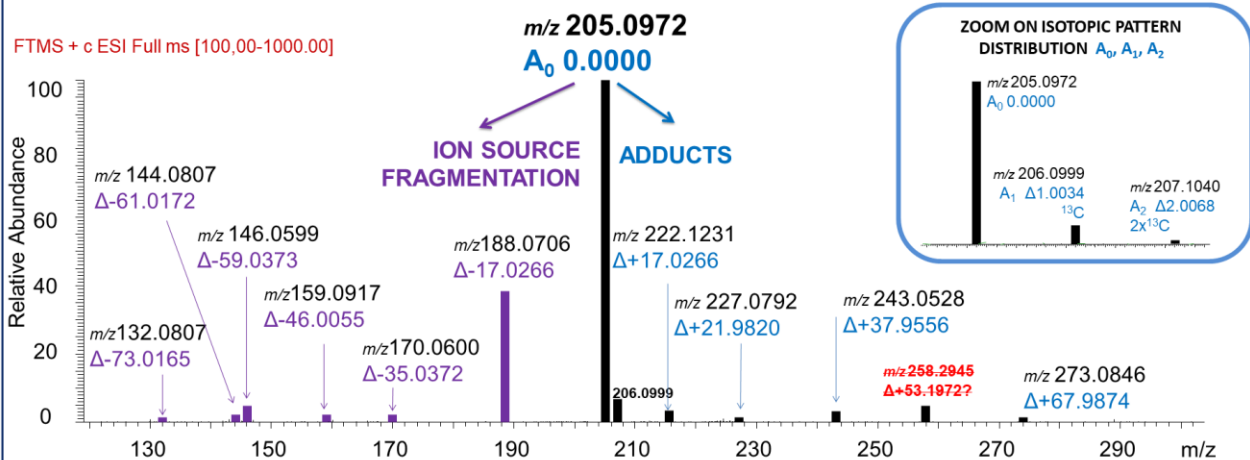


Box S9.2. Overview of identification steps in LC-MS and GC-MS.

carefully, as their masses might match $[M+H]^+$ and $[M-H]^-$ species from other metabolites. This might occur when metabolites structures only differ in labile chemical moieties and chemical substructures are shared with other metabolites (i.e. tryptophan and hydroxytryptophan or tryptamine and serotonin). In summary, the presence of two or more adducts together with neutral losses can provide very useful information and aid in the appropriate selection of the pseudomolecular ion by comparing their monoisotopic masses and calculating an experimental mass difference between two features. This experimental mass difference is then compared with the theoretical mass difference between two known adducts, i.e. a sodiated molecular ion will have a mass difference of 21.9819 Da in comparison with the protonated molecular ion, while the

loss of a water molecule will create a 18.1057 Da mass difference. If the experimental mass difference falls within a user specified window (ppm error) of the theoretical one, then these two features can be annotated and the neutral mass using the same rules can be calculated. The relative intensity among adduct peaks as well as the amount of formed adducts will vary from one metabolite to another, as well as for the same metabolite depending on experimental conditions. Another source of useful information comprises the isotopic pattern distribution, which is nicely recorded in high resolution full scan spectra. Depending on the mass resolution that is achieved with the instrument used, isotopes of carbon, sulphur, and nitrogen may be verified. **Box S9.3** presents key points related to hypothesis generation.

LC-MS ANNOTATION – ASSIGNMENT OF ELEMENTAL COMPOSITION TO *m/z* FEATURE



Common fragments and losses

| | |
|---------|--|
| 4.9554 | Na ⁺ ↔ NH ₄ ⁺ , salt adduct |
| 17.0265 | neutral ammonium loss |
| 17.0265 | NH ₄ ⁺ ↔ H ⁺ , salt adduct |
| 18.0105 | ± H ₂ O, water addition/loss |
| 21.9819 | Na ⁺ ↔ H ⁺ , salt adduct |
| 37.9558 | K ⁺ ↔ H ⁺ , salt adduct |
| 67.9874 | NaCOOH ⁺ ↔ H ⁺ , salt adduct |
| 43.9898 | ± [CO ₂] adduct or loss |
| 176.032 | - [glucuronic acid] |
| 79.9568 | - [SO ₃] |

For more adducts/fragments see Keller et al. [78]

| | | | |
|----------|---|-----------|--|
| 20.92933 | K ⁺ ↔ NH ₄ ⁺ , salt adduct | 42.01057 | ± COCH ₂ |
| 21.98194 | Na ⁺ ↔ H ⁺ , salt adduct | 42.04695 | ± C ₂ H ₅ , propylation |
| 24.99525 | CN ↔ H, nitrile compounds | 97.96738 | ± H ₂ SO ₄ , sulfur compounds |
| 27.01090 | ± HCN, nitrile compounds | 97.97690 | ± H ₃ PO ₄ , phosphorous compounds |
| 27.99492 | ± CO | 100.90140 | I ↔ CN, halogen exchange with cyano group |
| 28.00615 | - 2N, nitrogen loss, e.g. azido compounds (N ₂) | 109.90173 | I ↔ OH, halogen exchange with hydroxy group (typically -I + OH) |
| 28.03130 | ± C ₂ H ₄ , natural alkane chains such as fatty acids | 125.89665 | I ↔ H, halogen exchange |
| 29.97418 | NO ₂ ↔ NH ₂ , nitro compounds | 146.05791 | ± [deoxy-hexose-H ₂ O, C ₆ O ₄ H ₁₀], e.g. fucose |
| 29.99799 | -NO, nitroso compounds | 164.06848 | ± [deoxy-hexose-H ₂ O, C ₆ O ₅ H ₁₂], e.g. fucose |
| 31.97207 | ± S, sulfur compounds | 180.06339 | ± [hexose, C ₆ O ₆ H ₁₂], e.g. glucose, galactose, mannose, |
| 31.98983 | ± 2O, oxygen loss | 194.04266 | ± [glucuronic acid, C ₆ O ₇ H ₁₀] |
| 33.02146 | - NH ₂ OH, loss from hydroxamic acids | | |
| 33.96103 | Cl ↔ H, halogen exchange | | |
| 33.98772 | ± H ₂ S, sulfur compounds | | |
| 37.95588 | K ⁺ ↔ H ⁺ , salt adduct | | |
| 37.98916 | 2CN ↔ 2COOH, nitrile compounds | | |



Introduce *m/z* feature into tool for assignment of elemental composition. Example with web free tool:
predret.org/tools/mass-decomposition

Basic settings

Select type of ion

Positive Negative Positive Neutral

[M-H]- [M+H]+ [M+Na]+

[M-2H+Na]- [M+K]+

[M-2H+K]- [M+K]+

[M+Cl]-

[M-H+HCOOH]-

[M-H+HCOONa]-

***m/z* to decompose:**
205.0972

ppm error:
8

DECOMPOSE MASS

Advanced settings

ELEMENTS (click to expand)

Minimum components:
CO

Maximum components:
C40H100N6O12P2S2

FILTERS (click to expand)

DBE limits: -10 to 20

H/C ratio: 0.1 to 8

N/C ratio: 0 to 6

O/C ratio: 0 to 5

S/C ratio: 0.1 to 5

Formula hypotheses

Click columns to simulate the isotopic pattern.

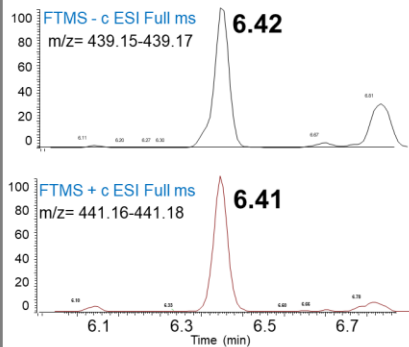
| Formula | Nitrogen rule | DBE | Calc. <i>m/z</i> | ppm | Ion | Adduct formula |
|------------|---------------|-----|------------------|--------|--------------------|-------------------------|
| C11H12N2O2 | Valid | 7 | 205.0972 | -0.225 | [M+H] ⁺ | C11H13N2O2 ⁺ |
| C5H13N6P | valid | 3 | 205.0961 | -5.28 | [M+H] ⁺ | C5H13N6P ⁺ |
| C9H17O3P | valid | 2 | 205.0988 | 7.88 | [M+H] ⁺ | C9H17O3P ⁺ |

Box S9.3. LC-MS annotation - part 1

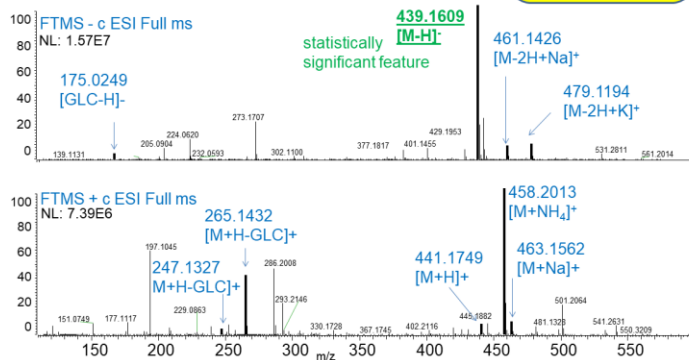
LC-MS ANNOTATION – ANALYSIS OF MS/MS SPECTRA



I) Simultaneous analysis of TIC in both ionization modes



II) Comparison of Full Scan spectra in both ionization modes



III) Comparison of MS/MS spectra in positive and negative ionization modes

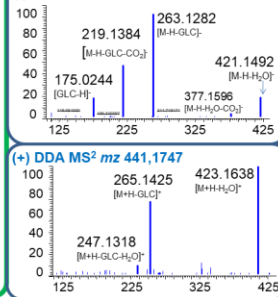
Information derived from Full Scan spectra:

- ✓ Elemental composition: $C_{21}H_{28}O_{10}$
- ✓ Presence of glucuronide group: characteristic ion 175.0242 (in negative mode), confirmed by neutral loss of 176.0320 moiety visible in positive ionization mode.
- ✓ Remaining aglicone moiety: $C_{15}H_{20}O_4$

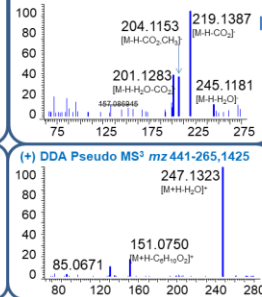
MS/MS spectra:

- Assign to each fragment elemental composition and propose plausible neutral loss of group or moiety

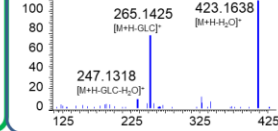
(-) DDA MS² m/z 439,1609



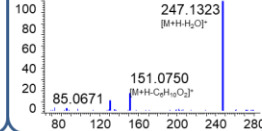
(-) Re-injection MS² m/z 439-263,1281



(+) DDA MS² m/z 441,1747



(+) DDA Pseudo MS² m/z 441-265,1425



Re-inject sample, if amount of information from DDA MS/MS spectra is not sufficient

IV) Comparison of MS/MS spectra with online database

mzCloud Search: 3 Results:

Verify number of fragments matching correctly between library and query, as well as their mass ppm errors and intensities

Ionization mode
 Positive Negative

Search Type
 Comp. Identification (Search in MS)
 Substr. Identification (Search in MS²)

Search In
 Filtered Spectra
 Recalibrated Spectra
 Average Spectra
 Composite Spectra

Algorithm
 HighChem High Res
 Opt.Dot. Product
 NIST (modified)

Spectrum Query

Precursor m/z
 None m/z 263,1280 ± 0.05

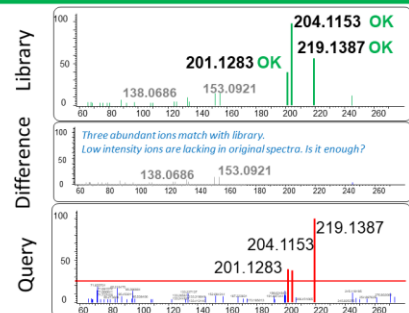
Accuracy (half width)
 From source Manual

Query Spectrum Noise Filter
 Number of highest peaks
 Intensity Threshold (%) 25,5

(±)-Abscisic acid
Match 77.9

Prostaglandin A1 ethyl ester
Match 23.9

7(S),17(S)-Dihydroxy-8(E),10(Z),13(Z),15(Z),19(Z),docosapenta-enoic acid
Match 6.8



Search for matching in both ionization modes, for as much ions/substructures as possible.

Introduce into the online spectral database the **smallest substructure, or moiety after conjugation loss** with good quality MS/MS spectra (here m/z 263.1281(neg) or m/z 247.1323(pos)). It is easier to find good substructure matching for small moieties than for whole unknown structure.

Search in more than one spectral database!



EXAMPLES OF TOOLS AND WEB SITES FOR ANNOTATIONS

- MassBank <http://www.massbank.jp/index.html>
- MetAssign-mzMatch <http://mzmatch.sourceforge.net/index.php>
- MetFrag <http://c-ruttikies.github.io/MetFrag>
- FingerID <https://github.com/icdishb/fingerid>
- MyCompoundID http://mycompoundid.org/mycompoundid_IsoMS
- MetFrag <https://msbi.ipb-halle.de/MetFrag/>
- MetFusion <https://msbi.ipb-halle.de/MetFusion/>
- CDM-ID <http://cfmid.wishartlab.com/>
- CSI:Finger ID <https://www.csi-fingerid.uni-jena.de/>

See more on: Spicer et al. [94]

Lynn et al.^[74] and Domingo-Almenara et al.^[75] summarized the ions to consider when interpreting MS data:

- Adducts: are formed by the addition of atoms or molecules to a metabolite. The proton adduct of $[M + H]^+$ in the positive ion mode and $[M - H]^-$ in the negative ion mode are most frequently observed in electrospray ionization. **Table S9.3** summarizes some of them.^[76]

Table S9.3. Selection of the most common adducts observed in electrospray ionization. The adducts with masses in italics have been taken from Keller et al.^[76]

| Mass difference | Origin |
|------------------|----------------------------------|
| Positive adducts | |
| 1.007276 | H |
| <i>18.01057</i> | H ₂ O, water clusters |
| 18.033823 | NH ₄ |
| 22.989218 | Na |
| 33.033489 | CH ₃ OH+H |
| 55.015431 | CH ₃ OH + Na |
| 38.963158 | K |
| 42.033823 | CH ₃ CN+H |
| 64.015765 | CH ₃ CN+Na |
| Negative adducts | |
| 1.007276 | H |
| <i>18.01057</i> | H ₂ O, water clusters |
| 44.998201 | HCOO-H |
| 59.013851 | CH ₃ COO-H |
| 112.985586 | CF ₃ COOH-H |

- Fragments: some weak bonds in a metabolite can be broken during the ionization process, producing a series of so-called in-source fragment peaks, observable in the full scan MS spectra. The main difficulty in mass spectrometry is that a mass spectrum obtained from different mass analysers could be slightly different because of the fragmentation during the ionization process. Some fragment peaks, which are called characteristic fragments, are unique to a certain metabolite class and may assist in metabolite identification.
- Isotopes: the m/z difference between the isotopic peaks of a metabolite reveals its charge state, whereas the relative abundances disclose its possible atom composition and the existence of coelution.
- Multimers: When the sample concentration is high, some metabolites can form multimers, mostly dimers and trimers. The abundance of the multimer is usually much lower than that of the monomer.

Annotation can be performed manually or with algorithms such as CAMERA (Collection of Algorithms for METabolite pRofile Annotation). This R-package tool is available on

Galaxy Web platform (Workflow4metabolomics.org). Its primary purpose is the annotation and evaluation of LC-MS data and includes algorithms for annotation of isotope peaks, adducts, and fragments in peak lists. Very recently, Domingo-Almenara et al.^[75] discussing the advantages and pitfalls of each of the state-of-the-art strategies in computational annotation including: i) peak grouping or full scan (MS1) pseudo-spectra extraction (i.e. clustering all mass spectral signals stemming from a single metabolite); and ii) annotation using ion adduction and mass distance among ion peaks and others.

When a monoisotopic ion is revealed, the elemental composition can be determined with a molecular formula generator. Usually a certain mass range must be entered to generate formulas in this range; also elements, which should be included or excluded need to be listed. The main problem with formula generators is that they produce too many wrong candidates. Chemometric rules must be applied to constrict the number of molecular formulas. Different approaches have been proposed, such as the 'seven golden rules' by Kind et al.^[77] These seven heuristic rules enable an automatic exclusion of molecular formulas, which are either wrong or which contain an unlikely high or low number of elements. Joerg Hau (<http://jhau.maliwi.de/sci/ms.html>) developed HiRes generator based on these rules. Another very useful, freely accessible, and user-friendly elemental composition generator is found online in the PredRet platform^[78] (predret.org/tools/mass-decomposition). The number of possible formulas can also be restricted to those existing in biology: either natural products (present in the CRC Dictionary of Natural Products (DNP database: <http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml>) or endogenous compounds (for example, HMDB, which provides information about the biological matrix in which the compound can be retrieved such as urine, blood, tissue, and feces).

When a unique molecular formula is established, annotation can be undertaken by matching it against a database. As already stated earlier, searches are usually not made within a single resource. The first step is therefore to query an in-house specific spectral library containing spectra of authentic standards analyzed on the same instrument. However, if no hit is returned, other libraries should be interrogated. This step can be performed on the Galaxy web platform W4M. This platform allows the creation of an automatic workflow allowing to query several databases (Mass Bank, HMDB, KEGG,...) and the automatic combining of results. At this point, hundreds of hypotheses can be generated and the challenge will then consist in the selection of a few most plausible ones. If a hypothesis is not retrieved, more general compound databases, such as ChEBI or PubChem, can be queried. *In silico* prediction tools can also be used. Several tools (Meteor Nexus, Metabolizer) and, more recently, databases such as MyCompoundID^[79] and IIMDB^[80] can generate phase I and phase II predicted metabolites for exogenous compounds.

9.4.2. Hypothesis Selection

LC coupled to tandem mass spectrometry experiments allow to access structural information. Fragmentation experiments are therefore performed in order to select or validate generated hypotheses. CID and HCD are the most common ion activation techniques used in metabolomics. However, the resulting low-energy CID spectra are not very reproducible and depend on: i) the instrument

characteristics (e.g. quadrupole *versus* ion trap or TOF mass spectrometers...); ii) the lens potentials; iii) the collision energy value; and iv) the target gas pressure. Building databases for CID in MS/MS experiments is consequently difficult and requires a very cautious control of the experimental parameters.^[81]

Different instruments and different spectrometric methods may be used for generating fragmentation spectra. This topic is presented in detail in chapter 6. Acquisition of MS/MS fragmentation spectra with low or high resolution is a fundamental and a crucial step for the structural elucidation and appropriate annotation of molecules. As mentioned in the previous section, MS/MS spectra can be acquired at low or high resolution. However, high resolution MS/MS spectra enable a more reliable description of fragmentation mechanisms. Manual MS/MS data interpretation requires a high degree of expertise and is very time consuming, while automatic annotations are not yet fully trustworthy. Therefore, the next step in annotation is to compare spectra of potential new markers with in-house and/or on-line mass spectrometry libraries.

Among the most popular libraries are HMDB, Metlin, MassBank, mzCloud and LipidMaps containing CID and HCD spectra. Original spectra from the compound under investigation can be directly copied to databases (mzCloud) or precursor and fragment ions can be manually introduced on the website together with their intensities. The output of a search is usually given in a table with a score value describing the extent of matching between two spectra. The reproducibility of mass spectra of whole compounds or their substructures among Orbitrap equipment is relatively high. Thus, for experiments in Orbitrap, searching the mzCloud database should be the first choice. Spectral comparison between ToF and Orbitrap instruments, or even between different types of ToF instruments, can be less reproducible and require more intensive searches. For this reason, MS/MS scoring algorithms are still challenging. Different software tools are currently available: NIST MS Search (<http://chemdata.nist.gov>), MS-DIAL,^[82] Mass Frontier (Thermo Fisher Scientific, Waltham, USA), SmileMS (<http://www.genebio.com>), Mass++,^[83] and XCMS2.^[84] The performance of these tools in performing MS/MS searches still needs validation, in particular considering false positives and *in silico* databases.^[85]

To facilitate the annotation process, it is advisable to investigate MS and MS/MS data in the positive and negative ionization mode simultaneously. The information acquired by both modes is complementary and provides important information on the nature of the compound of interest. An example of such complementary information derived from both ionization modes is given in **Box S9.4**. In the negative ion mode, the statistically significant pseudomolecular ion 439.1609 [M-H]⁻ dominates the spectrum, and the deprotonated glucuronide ion *m/z* 175.0249 is also visible, what suggests that the compound of interest is a conjugated metabolite, although no aglycone is present in the spectra. Also, although a protonated aglycone ion is visible in positive ionization mode, no pseudo molecular ion [M+H]⁺ is observed, in addition to ammonium and sodium adducts. In case of conjugated metabolites (with glucuronide, sulphate, or glycine group), online searching should start with deconjugated moieties, as these molecules provide a better matching than the conjugated metabolite. In case that MS/MS information from an automatic DDA/DIA experiment is not sufficient, the sample(s) should be re-injected using SIM, or other dedicated MS/MS methods selecting ion(s) of

interest. When neither online nor in-house libraries support the identification of an unknown compound, a manual investigation should be performed, which starts with the assignment of each fragment ion to an elemental composition and the identification of possible plausible moieties lost from the structure. An interpretation of the fragmentation pattern requires deep knowledge, not only of mass spectrometry, but also of organic and physical chemistry. However, access to this knowledge is supported by the availability of mass spectra of common compound classes as well as of the main mechanisms of fragmentation.^[86, 87]

Different computational approaches have been developed to automate the annotation process and propose molecular structures. Böcker and Rasche^[88] presented a method for annotating tandem MS data using a hypothetical fragmentation tree. Tree nodes are annotated with the molecular formula of the fragments, the edges representing (neutral or radical) losses. More recently, novel *in silico* fragmentation and prediction algorithms were developed for the identification of unknown features. In particular, Rasche et al.^[89] developed a method based on local tree alignments using the parts of trees in which similar fragmentation cascades occur. Expert databases can also be used for filtering hypotheses by taking the biological plausibility into account. In particular, HMDB, FooDB and PhytoHUB have been designed for nutritional metabolomics and integrate spectral data, information on dietary origin and compound metabolism in humans.

At the end of this step, either commercial standards are available for the unique selected hypotheses, and identification can be validated quite easily; or standards are lacking and identifications will remain putative. It is important to note that chemical or *in vitro* enzymatic syntheses can be interesting approaches for producing new standards. The synthesis or isolation of a missing compound can be requested on the virtual board of FoodComEx, the online collaborative platform for sharing of food-derived standards (<http://foodcomex.org/>).

9.4.3. Complementary Experiments

Complementary experiments are necessary for *de novo* structural elucidation, in particular for obtaining additional evidence on the structural plausibility of level 2 metabolites or providing information on the identity of level 3 metabolites. To this end, the number of exchangeable hydrogen atoms (i.e. attached to O, N, and S atoms) can be determined using hydrogen/deuterium exchange in solution for elucidating mass fragmentation mechanisms.^[90] Besides, other analytical tools (NMR, GC/MS, UV,...) can be used to produce complementary structural information for identifying unknown features. However, this approach often requires the purification or isolation of these compounds in reasonable quantities using preparative HPLC, LC-NMR, LC-SPE-NMR, or LC fraction collection.^[91]

9.5. Identification of Compounds in NMR-based Metabolome Studies

One of the most challenging steps in NMR analysis is the accurate identification of metabolites and quantification of their abundances, especially due to the fact that some spectral regions appear very crowded and are characterized by a heavy peak overlap. Different tools are now available for these tasks. These include the use of performant and

popular online and offline databases, the conductance of specific 2D experiments (to better characterize unknown metabolites), and the utilization of standards (to identify or quantify metabolites).

9.5.1. Database

Many online databases are currently available for the identification of metabolites based on the identification of specific experimental NMR signals. For example, HMDB (www.hmdb.ca) and BMRB (www.bmrwisc.edu) list known chemical shifts that can be matched with the queried signals. In cases in which the objective is to find if a specific molecule is present in a particular sample and the relative spectrum is not available in existing libraries, other online tools can be helpful. Websites such as NMRDB (nmrdb.org) allow the drawing of the molecular structure of a metabolite and the uploading of its molfile in order to predict its NMR spectrum. Spectral characteristics of the molecules of interest, such as their chemical shift, multiplicity, and J-coupling, can thus be derived.

Compared to online solutions, offline software and databases have been more recently developed. Files containing the experimental NMR spectra can be downloaded and matched with the spectral files of the molecule contained in the software's databases (**Box S9.5**):

- *Chenomx* is an offline library containing a profiler that allows the identification and quantification of a large number of metabolites. The *Chenomx* library consists of hundreds of molecules, each molecule being characterized by a range of parameters (biological function, clinical relevance, molecular structure, and usual concentrations in biofluids). The *Chenomx* profiler allows the automatic searching of specific compounds using their names or fits chemical shifts to the best possible compounds. The combination of many different signals from different molecules is also possible, in order to match the analyzed spectrum and find out the composition of a sample.
- *Bayesil* (www.bayesil.ca) can be used for metabolite identification and quantitation in human blood samples.^[92] This fully-automated web-based system can perform all pre-processing steps on spectra, deconvolute them using a reference library containing more than 60 metabolites, and rapidly determine the identity (90% correct identification) and concentration (10% quantification error)

Finally, software programs that are specific for the NMR instrument, such as *AssureNMR* for Bruker spectrometers, combine tools for statistical data analysis with the identification and quantitation of molecules (<https://www.bruker.com/products/mr/nmr/nmr-software/nmr-software/assurenmr/overview.html>).

9.5.2. Spiking

Another way of assessing the true identity of a signal is through spiking experiments. These experiments consist in comparing the NMR spectrum of a sample with its spectrum

after the spiking of a known molecule that could be responsible for a signal. In this way, it is possible to validate or discard spectral assignments. Spiking can also be used to calculate the concentration of specific metabolites. In this case, a known amount of an internal calibration standard is used to spike the sample. This method, although very accurate, is flawed with disadvantages. In particular, spiking irreversibly contaminates samples, so that these samples cannot be recovered. When using spiking for absolute quantitative analysis, great precision and care must be given to the mixture preparation. Moreover, as for all the quantitative purposes, the spin-lattice relaxation time T_1 of all protons in the sample must be known to choose the proper experimental relaxation time. In addition, the reference compound needs to be suitable for the experiment and must be i) soluble and inert; ii) stable in time; iii) preferably characterized by a shorter T_1 compared to the investigated metabolites; and iv) demonstrate a minimal signal overlap with both the solvent and the analyzed metabolites.

9.5.3. Two-Dimensional Experiments

2D-NMR experiments can be also used to identify unknown signals, in particular since 2D-NMR sequences provide information which is otherwise not available from 1D spectra. In classical 2D experiments, a pulse sequence excites the nuclei in the samples with two or more pulses and the FID is recorded. The acquisition is then carried out many times, in order to increase evolution time (t_1) between pulse groups. The acquisition time that characterizes the 2D NMR experiment is denominated as t_2 . Therefore, 2D spectra will be characterized by a stack of 1D spectra, differing by a slight variation in the t_1 value. All parameters are kept constant in each successive experiment, apart from the pulse phase. After Fourier transformation of the first time domain and of the interferogram on the second dimension, a map of spin-spin correlations is created, with the axes of this map being the evolution frequency (f_1) and the acquisition frequency (f_2). If a signal's frequency has changed during t_1 , f_1 will differ from f_2 . In order to more clearly represent this approach, the 2D spectrum is usually plotted with their 1D projection on the axes.

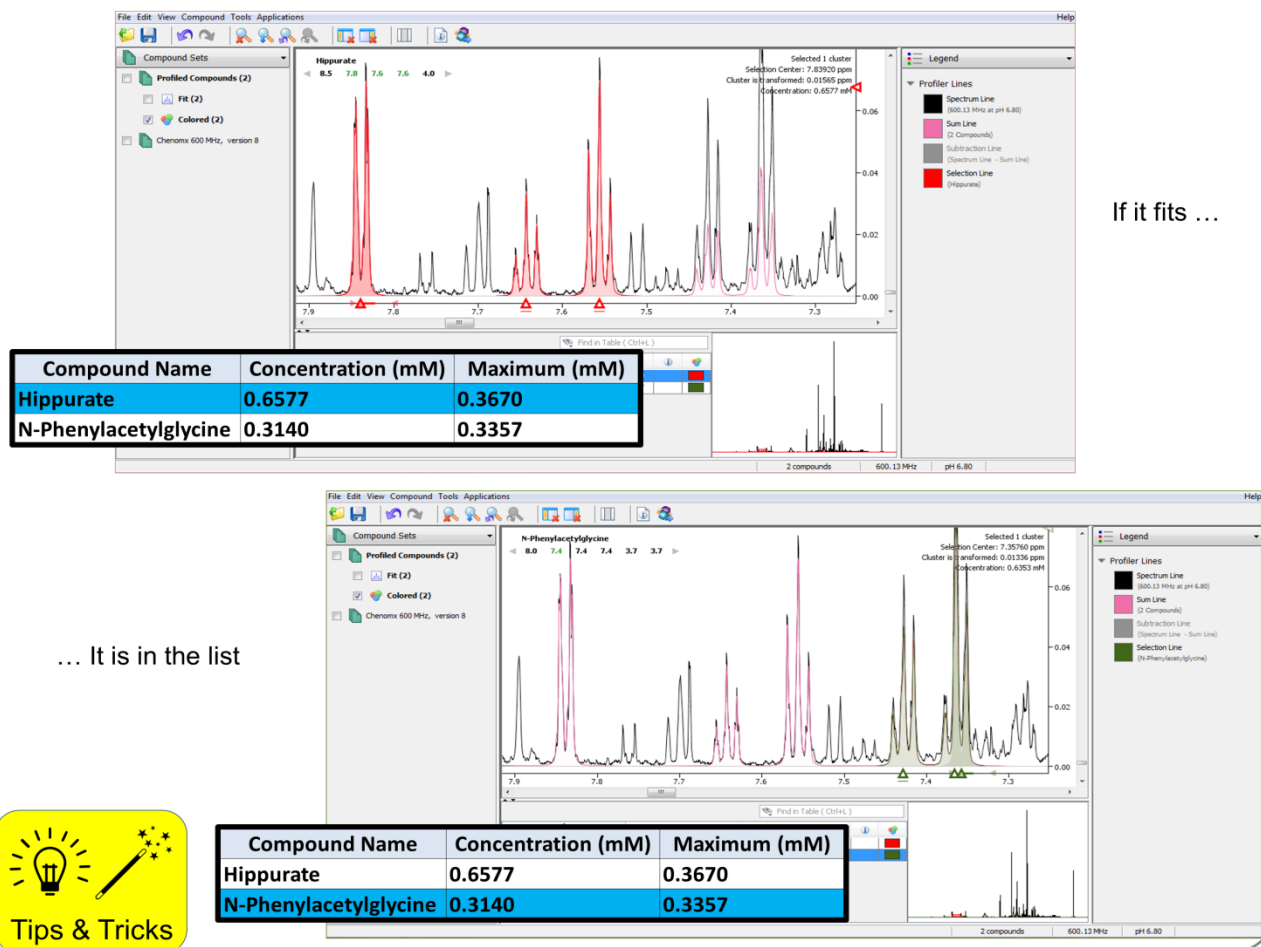
A two-dimensional spectrum will be composed of two types of signals:

- Cross-peaks, giving information on the correlation between two different resonances originated by two nuclei at short distance or connected through a bond.
- Diagonal peaks, representing the frequency of the signals in the original 1D spectrum.

A pseudo-2D NMR spectrum can also be generated through the Statistical Total Correlation Spectroscopy (STOCSY) method, which was implemented specifically for metabolite identification in metabolomic investigations. The multicollinearity of the intensities in a spectral dataset is used to generate this pseudo-2D spectrum, displaying the correlation between peak intensities.^[93]

METABOLITES IDENTIFICATION IN NMR

The identification of metabolites in 1H-NMR spectra can be made by comparing with NMR spectra of reference compounds available in dedicated commercial softwares as Chenomx (<http://www.chenomx.com>)



Box S9.5. NMR annotation.

References

- [1] M. Baker, *Nat. Methods* **2011**, 8, 117.
- [2] D. J. Creek, J. Darren, W. B. Dunn, O. Fiehn, J. L. Griffin, R. D. Hall, Z. Lei, R. Mistrik, *Metabolomics* **2014**, 10, 350.
- [3] L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden, M. R. Viant, *Metabolomics* **2007**, 3, 211.
- [4] M. R. Viant, I. J. Kurland, M. R. Jones, W. B. Dunn, *Curr. Opin. Chem. Biol.* **2017**, 36, 64.
- [5] N. Strehmel, J. Hummel, A. Erban, K. Strassburg, J. Kopka, *J. Chromatogr. B* **2008**, 871, 182.
- [6] L. W. Sumner, Z. Lei, B. J. Nikolau, K. Saito, U. Roessner, R. Trengove, *Metabolomics* **2014**, 10, 1047_1049.
- [7] R. M. Salek, C. Steinbeck, M. R. Viant, R. Goodacre, W. B. Dunn, *GigaScience* **2013**, 2, 13.
- [8] M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek, O. Yanes, *TrAC Trends Anal. Chem.* **2016**, 78, 23.
- [9] Z. Lacroix, T. Critchlow, *Bioinformatics: Managing Scientific Data*, Morgan Kaufmann Publishers Inc, **2003**.
- [10] H. E. Pence, A. Williams, *J. Chem. Educ.* **2010**, 87, 1123.
- [11] O. Fiehn, D. K. Barupal, T. Kind, *J. Biol. Chem.* **2011**, 286, 23637.
- [12] J. Hastings, P. De Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, C. Steinbeck, *Nucleic Acids Res.* **2013**, 41, D456.
- [13] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, *J. Mass Spectrom.* **2010**, 45, 703.
- [14] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, G. Siuzdak, *Ther. Drug Mon.* **2005**, 27, 747.
- [15] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, D. Steinhauser, *Bioinformatics* **2005**, 21, 1635.
- [16] T. Kind, G. Wohlgemuth, D. Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, O. Fiehn, *Anal. Chem* **2009**, 81, 10038.
- [17] C. A. Valdez, R. N. Leif, S. Hok, A. Alcaraz, *J. Mass Spectrom.* **2018**, 53, 419.
- [18] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vazquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, A. Scalbert, *Nucleic Acids Res.* **2018**, 46, D608.
- [19] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. De Matos, M. Rijnbeek, T. Mahendrakar, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. Gonzalez-Beltran, S. A. Sansone, J. L. Griffin, C. Steinbeck, *Nucleic Acids Res.* **2013**, 41, D781.

- [20] M. Van Rijswijk, C. Beirnaert, C. Caron, M. Cascante, V. Dominguez, W. B. Dunn, T. M. D. Ebbels, F. Giacomoni, A. Gonzalez-Beltran, T. Hankemeier, K. Haug, J. L. Izquierdo-Garcia, R. C. Jimenez, F. Jourdan, N. Kale, M. I. Klapa, O. Kohlbacher, K. Koort, K. Kultima, G. Le Corguille, N. K. Moschonas, S. Neumann, C. O'Donovan, M. Reczko, P. Rocca-Serra, A. Rosato, R. M. Salek, S. A. Sansone, V. Satagopam, D. Schober, R. Shimmo, R. A. Spicer, O. Spjuth, E. A. Thevenot, M. R. Viant, R. J. M. Weber, E. L. Willighagen, G. Zanetti, C. Steinbeck, *F1000Res* **2017**, 6, 1649.
- [21] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, 44, D1202.
- [22] A. G. Sharkey, R. A. Friedel, S. H. Langer, *Anal. Chem.* **1957**, 29, 770.
- [23] D. C. Dejongh, T. Radford, J. D. Hribar, S. Hanessian, M. Bieber, G. Dawson, C. C. Sweeley, *J. Am. Chem. Soc.* **1969**, 91, 1728.
- [24] M. Zinbo, W. R. Sherman, *J. Am. Chem. Soc.* **1970**, 92, 2105.
- [25] R. A. Laine, C. C. Sweeley, *Anal. Biochem.* **1971**, 43, 533.
- [26] D. J. Harvey, M. G. Horning, *J. Chromatogr. A* **1973**, 76, 51.
- [27] R. A. Laine, C. C. Sweeley, *Carbohydr. Res.* **1973**, 27, 199.
- [28] P. L. Coduti, C. A. Bush, *Anal. Biochem.* **1977**, 78, 21.
- [29] Z. Füzfai, I. Boldizsár, I. Molnár-Perl, *J. Chromatogr. A* **2008**, 1177, 183.
- [30] G. Petersson, *Tetrahedron* **1969**, 25, 4437.
- [31] W. R. Sherman, N. C. Eilers, S. L. Goodwin, *Org. Mass Spectrom.* **1970**, 3, 829.
- [32] T. Niwa, N. Yamamoto, K. Maeda, K. Yamada, T. Ohki, M. Mori, *J. Chromatogr. B* **1983**, 277, 25.
- [33] K. Bergström, J. Güttler, R. Blomstrand, *Anal. Biochem.* **1970**, 34, 74.
- [34] F. P. Abramson, M. W. McCaman, R. E. McCaman, *Anal. Biochem.* **1974**, 57, 482.
- [35] J. Mařík, A. Čapek, J. Králíček, *J. Chromatogr. A* **1976**, 128, 1.
- [36] S. E. Hattox, R. C. Murphy, *Biol. Mass Spectrom.* **1978**, 5, 338.
- [37] I. Horman, R. Viani, *Organic Mass Spectrometry* **1971**, 5, 203.
- [38] G. Petersson, *Org. Mass Spectrom.* **1972**, 6, 565.
- [39] A. M. Lawson, R. A. Chalmers, R. W. E. Watts, *Biol. Mass Spectrom.* **1974**, 1, 199.
- [40] R. J. Horvat, S. D. Senter, *Org. Mass Spectrom.* **1983**, 18, 413.
- [41] B. M. Scholz-Böttcher, L. Ernst, H. G. Maier, *Liebigs Ann. Chem.* **1991**, 1991, 1029.
- [42] J. F. Rontani, C. Aubert, *Rapid Commun. Mass Spectrom.* **2004**, 18, 1889.
- [43] E. White, P. M. Krueger, J. A. McCloskey, *J. of Org. Chem.* **1972**, 37, 430.
- [44] A. E. Pierce, *Silylation of organic compounds: a technique for gas-phase analysis*, Rockford 1968, pp. 33-39.
- [45] J. M. Halket, V. G. Zaikin, *Eur. J. Mass Spectrom.* **2003**, 9, 1.
- [46] J. H. Gross, *Mass Spectrometry - a textbook*, Springer 2004.
- [47] S. Thurnhofer, W. Vetter, *J. Agric. Food Chem.* **2005**, 53, 8896.
- [48] P. Yin, H. Chen, X. Liu, Q. Wang, Y. Jiang, R. Pan, *Anal. Lett.* **2014**, 47, 1579.
- [49] D. J. Harvey, M. G. Horning, P. Vouros, *J. Chem. Soc. Chem. Comm.* **1970**, 898.
- [50] J. M. Halket, A. Przyborowska, S. E. Stein, W. G. Mallard, S. Down, R. A. Chalmers, *Rapid Commun. Mass Spectrom.* **1999**, 13, 279.
- [51] J. Gullberg, P. Jonsson, A. Nordström, M. Sjöström, T. Moritz, *Anal. Biochem.* **2004**, 331, 283.
- [52] N. Strehmel, J. Kopka, D. Scheel, C. Böttcher, *Metabolomics* **2014**, 10, 324.
- [53] A. D. K. Michaël Méret, B. Thomas Degenkolbe, A. Sabrina Kleessen, A. Zoran Nikoloski, A. Verena Tellstroem, C. M. Meret, D. Kopetzki, T. Degenkolbe, S. Kleessen, Z. Nikoloski, V. Tellstroem, A. Barsch, J. Kopka, M. Antonietti, L. Willmitzera, *RSC Adv.* **2014**, 4, 16777.
- [54] J. Kopka, S. Schmidt, F. Dethloff, N. Pade, S. Berendt, M. Schottkowski, N. Martin, U. Duhring, E. Kuchmina, H. Enke, D. Kramer, A. Wilde, M. Hagemann, A. Friedrich, *Biotechnol. biofuels* **2017**, 10, 56.
- [55] E. Kováts, *Helv. Chim. Acta* **1958**, 41, 1915.
- [56] H. Van Den Dool, P. D. Kratz, *J. Chromatogr. A* **1963**, 11, 463.
- [57] G. Castello, *J. Chromatogr. A* **1999**, 842, 51.
- [58] B. D'acampora Zellner, C. Bicchì, P. Dugo, P. Rubiolo, G. Dugo, L. Mondello, *Flavour Frag. J.* **2008**, 23, 297.
- [59] M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek, O. Yanes, *TrAC Trends Anal. Chem.* **2016**, 78, 23.
- [60] R. J. Western, P. J. Marriott, *J. Chromatogr. A* **2003**, 1019, 3.
- [61] S. Bieri, P. J. Marriott, *Anal. Chem.* **2008**, 80, 760.
- [62] M. Jiang, C. Kulsing, Y. Nolvachai, P. J. Marriott, *Anal. Chem.* **2015**, 87, 5753.
- [63] E. M. Brouwer-Brolsma, L. Brennan, C. A. Drevon, H. Van Kranen, C. Manach, L. O. Dragsted, H. M. Roche, C. Andres-Lacueva, S. J. L. Bakker, J. Bouwman, F. Capozzi, S. De Saeger, T. E. Gundersen, M. Kolehmainen, S. E. Kulling, R. Landberg, J. Linseisen, F. Mattivi, R. P. Mensink, C. Scaccini, T. Skurk, I. Tetens, G. Vergeres, D. S. Wishart, A. Scalbert, E. J. M. Feskens, *P. Nutr. Soc.* **2017**, 76, 619.
- [64] D. S. Wishart, *Bioanalysis* **2009**, 1, 1579.
- [65] K. Scheubert, F. Hufsky, S. Böcker, *J. Cheminformatics* **2013**, 5, 12.
- [66] T. Kind, O. Fiehn, *BMC Bioinformatics* **2007**, 8, 105.
- [67] K. Varmuza, W. Werther, *J. Chem. Inf. Comp. Sci.* **1996**, 36, 323.
- [68] J. Hummel, N. Strehmel, J. Selbig, D. Walther, J. Kopka, *Metabolomics* **2010**, 6, 322.
- [69] A. L. Kerber, R. M. Meringer, K. Varmuza, *Adv. Mass Spectrom.* **2001**, 15, 939.
- [70] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, *BMC Bioinformatics* **2010**, 11, 148.
- [71] E. L. Schymanski, C. M. J. Gallampois, M. Krauss, M. Meringer, S. Neumann, T. Schulze, S. Wolf, W. Brack, *Anal. Chem.* **2012**, 84, 3287.
- [72] P. T. Kasper, M. Rojas-Chertó, R. Mistrik, T. Reijmers, T. Hankemeier, R. J. Vreeken, *Rapid Commun. Mass Spectrom.* **2012**, 26, 2275.
- [73] F. Hufsky, M. Rempt, F. Rasche, G. Pohnert, S. Böcker, *Anal. Chim. Acta* **2012**, 739, 67.
- [74] K. S. E. A. Lynn, *Anal. Chem.* **2015**, 87, 2143.
- [75] X. Domingo-Almenara, J. R. Montenegro-Burke, H. P. Benton, G. Siuzdak, *Anal. Chem.* **2018**, 90, 480.
- [76] B. O. Keller, J. Sui, A. B. Young, R. M. Whittall, *Anal. Chim. Acta* **2008**, 627, 71.
- [77] T. Kind, O. Fiehn, *BMC Bioinformatics* **2007**, 8, 105.
- [78] J. Stanstrup, S. Neumann, U. Vrhovsek, *Anal. Chem.* **2015**, 87, 9421.
- [79] L. Li, R. Li, J. Zhou, A. Zuniga, A. E. Stanislaus, Y. Wu, T. Huan, J. Zheng, Y. Shi, D. S. Wishart, G. Lin, *Anal. Chem.* **2013**, 85, 3401.
- [80] L. C. Menikarachchi, D. W. Hill, M. A. Hamdalla, Mandoiu, Ii, D. F. Grant, *J. Chem. Inf. Model.* **2013**, 53, 2483.
- [81] E. Werner, J. F. Heilier, C. Ducruix, E. Ezan, C. Junot, J. C. Tabet, *J. Chromatogr. B* **2008**, 871, 143.
- [82] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. Vanderghenst, O. Fiehn, M. Arita, *Nat. Methods* **2015**, 12, 523.
- [83] S. Tanaka, Y. Fujita, H. E. Parry, A. C. Yoshizawa, K. Morimoto, M. Murase, Y. Yamada, J. Yao, S. I. Utsunomiya, S. Kajihara, M. Fukuda, M. Ikawa, T. Tabata, K. Takahashi, K. Aoshima, Y. Nihei, T. Nishioka, Y. Oda, K. Tanaka, *J. Proteome Res.* **2014**, 13, 3846.
- [84] H. P. Benton, D. M. Wong, S. A. Trauger, G. Siuzdak, *Anal. Chem.* **2008**, 80, 6382.
- [85] T. Kind, H. Tsugawa, T. Cajka, Y. Ma, Z. Lai, S. S. Mehta, G. Wohlgenuth, D. K. Barupal, M. R. Showalter, M. Arita, O. Fiehn, *Mass Spectrom. Rev.* **2017**, 37, 513.
- [86] F. W. McLafferty, F. Turecek, *Interpretation of mass spectra*, University Science Books, Mill Valley, USA **1993**.
- [87] E. Pretsch, P. Bühlmann, M. Badertscher, *Structure Determination of Organic Compounds*, Springer, Berlin **2009**.
- [88] S. Böcker, F. Rasche, *Bioinformatics* **2008**, 24, i49.
- [89] F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatos, S. Bocker, *Anal. Chem.* **2012**, 84, 3417.
- [90] D. Q. Liu, C. E. Hop, M. G. Beconi, A. Mao, S. Chiu, *Rapid Commun. Mass Spectrom.* **2001**, 15, 1832.
- [91] D. A. Dias, O. A. Jones, D. J. Beale, B. A. Boughton, D. Benheim, K. A. Kouremenos, J. L. Wolfender, D. S. Wishart, *Metabolites* **2016**, 6, 46.

- [92] S. Ravanbakhsh, P. Liu, T. C. Bjorndahl, R. Mandal, J. R. Grant, M. Wilson, R. Eisner, I. Sinelnikov, X. Hu, C. Luchinat, R. Greiner, D. S. Wishart, *PLoS One* **2015**, 10, e0124219.
- [93] O. Cloarec, M. E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes, J. Nicholson, *Anal. Chem.* **2005**, 77, 1282.
- [94] R. Spicer, R.M. Salek, P. Moreno, D. Canueto, C. Steinbeck, *Metabolomics* **2017**, 13, 106.