Metabolomics (2024) 20:103
https://doi.org/10.1007/s11306-024-02167-2

REVIEW ARTICLE

Navigating common pitfalls in metabolite identification and metabolomics bioinformatics

Elva María Novoa-del-Toro[1] · **Michael Witting**[2,3]

Received: 13 June 2024 / Accepted: 31 August 2024

Abstract

Background Metabolomics, the systematic analysis of small molecules in a given biological system, emerged as a powerful tool for different research questions. Newer, better, and faster methods have increased the coverage of metabolites that can be detected and identified in a shorter amount of time, generating highly dense datasets. While technology for metabolomics is still advancing, another rapidly growing field is metabolomics data analysis including metabolite identification. Within the next years, there will be a high demand for bioinformaticians and data scientists capable of analyzing metabolomics data as well as chemists capable of using in-silico tools for metabolite identification. However, metabolomics is often not included in bioinformatics curricula, nor does analytical chemistry address the challenges associated with advanced in-silico tools.

Aim of review In this educational review, we briefly summarize some key concepts and pitfalls we have encountered in a collaboration between a bioinformatician (originally not trained for metabolomics) and an analytical chemist. We identified that many misunderstandings arise from differences in knowledge about metabolite annotation and identification, and the proper use of bioinformatics approaches for these tasks. We hope that this article helps other bioinformaticians (as well as other scientists) entering the field of metabolomics bioinformatics, especially for metabolite identification, to quickly learn the necessary concepts for a successful collaboration with analytical chemists.

Key scientific concepts of review We summarize important concepts related to LC-MS/MS based non-targeted metabolomics and compare them with other data types bioinformaticians are potentially familiar with. Drawing these parallels will help foster the learning of key aspects of metabolomics.

Keywords Metabolomics · Bioinformatics · Mass Spectrometry · Metabolite Identification · Metabolite databases · Data Analysis · LC-MS/MS

1 Introduction

Metabolomics, the systematic measurement of small molecules (< 1500 Da) in a given biological system, represents the newest addition to omics technologies. This approach holds great promise as metabolism is closely linked to the observed phenotype, and metabolomics has shown a tremendous increase in applications over the last few years. Data obtained from the ever-improving analytical methods used in metabolomics are becoming increasingly complex and require the development of more sophisticated data analysis approaches.

Metabolomics works at the interface between biochemistry, analytical chemistry, and bioinformatics and chemoinformatics. Many novel tools for processing and analyzing metabolomics data are published yearly, and there is an increasing demand for scientists capable of understanding and using them. However, metabolites exhibit features distinct from other biological molecules such as DNA, RNA, and proteins. Nowadays, education and training in bioinformatics are mostly centered around these molecules and often overlook the requirements for analyzing small molecules.

✉ Michael Witting
michael.witting@helmholtz-munich.de

[1] Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP- Purpan, UPS, 180 chemin de Tournefeuille St-Martin-du-Touch, BP 3, Toulouse Cedex 31931, France

[2] Metabolomics and Proteomics Core, Helmholtz Zentrum München, 85764 Neuherberg, Germany

[3] Chair of Analytical Food Chemistry, TUM School of Life Sciences, Technical University of Munich, 85354 Freising-Weihenstephan, Germany

Published online: 21 September 2024

Different courses try to close this gap, but structured programs are still missing. Therefore, bioinformaticians aiming to enter the field are often overwhelmed by the differences in data structures, requirements, etc.… Individual examples and attempts to close these gaps are created, e.g., the publicly available script "Algorithmic Mass Spectrometry" by Sebastian Böcker (Böcker, 2019). As highlighted in the title of the 2015 review by Johnson et al., "Bioinformatics: The Next Frontier of Metabolomics"(Johnson et al., 2015).

In this article, we do not aim to cover all possible topics of bioinformatics in metabolomics or to comprehensively review available tools; for this, we would like to refer to the excellent reviews conducted by Misra and colleagues (Misra, 2021). Instead, we will discuss the issues and pitfalls we encountered while working together within the framework of the MetClassNet project from 2021 to 2023. In aiming to develop new network-based approaches for analyzing metabolomics data, we began working together and had to learn the scientific terminology of each other's disciplines. Here, we summarize some pitfalls and concepts that need to be addressed for successful collaboration. We hope that by giving some examples and advice, we can help other bioinformaticians enter the field.

2 What can go wrong? Can't be that hard…

Working with metabolomics data means exposure to different data formats, structures, and problems. We point out four explicit examples that we came across that confused our daily collaboration. Since our project focused on non-targeted metabolomics data obtained by Liquid Chromatography-Mass Spectrometry, including tandem mass spectrometry for generation of fragmentation patterns, (LC-MS/MS), we will also focus on this technique. We suggest several excellent articles and reviews for an overview of LC-MS/MS in metabolomics (Alseekh et al., 2021; Zhou et al., 2012). Here, we focus on a concise description of the data structure. Different software tools for processing LC-MS/MS raw data exist, e.g., xcms, MZmine, MS-DIAL, or commercial solutions from LC-MS/MS vendors or independent suppliers (Benton et al., 2008; Schmid et al., 2023; Tsugawa et al., 2020). Depending on the entry point into a project, a bioinformatician has to deal with raw data and its preprocessing, which includes steps like calibration of the mass-to-charge ratio (*m/z*) axis, peak detection, and chromatographic alignment. Joint work between the authors was based on a so-called feature table and corresponding fragmentation spectra obtained from different LC-MS/MS experiments. This data type was our project entry point and will be the starting point for the following discussion.

The feature table contains multiple columns: *m/z*, retention time (RT), and peak intensities or areas at a minimum. Still, it can also include additional columns, depending on the LC-MS/MS and processing software used, e.g., Collisional Cross Sections (CCS) or peak quality parameters used and exported by the different software tools. The *m/z* and RT pair is typically unique to a feature. Peak intensities or areas represent the quantification value. The higher this value, the higher the concentration of a metabolite. However, these values cannot be directly transferred to concentration values and need calibration to establish a relationship between the concentration and peak intensities/areas. Additionally, this calibration is unique to each metabolite.

The peak intensities/areas are typically used for biological analysis using uni- or multi-variate statistics. In the case of data-dependent (DDA) or data-independent acquisition (DIA) modes, features are often associated with fragmentation spectra (see Fig. 1), which are essential for identifying metabolites. It should be noted that not every feature will have a fragmentation spectrum and coverage for current LC-MS/MS instrumentation is somewhere between 30% and 60%. In several cases, laboratories might perform full-scan analysis, meaning only MS^1 data will be collected during profiling of samples and targeted MS^2 data will be collected afterwards for features found to be statistically significant.

To facilitate the explanations dedicated to bioinformaticians, we will try to compare metabolomics data to data obtained from, for example, RNA-seq or proteomics (if possible). Additionally, we would like to mention that while LC-MS/MS raw data often uses a standard open data format (.mzML), peak tables from different software tools might look very different. Recent initiatives promoting standardized tabular formats, such as .mzTab, need to be taken up by the software community but are highly welcome (Griss et al., 2014; Hoffmann et al., 2019).

3 One metabolite, many, many signals

LC-MS/MS data is complex. A single metabolite can produce multiple signals dependent on its chemical formula and structure. In the ion source of the MS, ions are formed from metabolites and transferred to the gas phase for analysis. Depending on whether the positive or negative ionization mode is used, cationic or ionic pseudo-molecular ions are generated. Certain metabolites will only ionize in one ionization mode, while others can be ionized in both modes. Main adducts are $[M+H]^+$ and $[M\text{-}H]^-$ yielded by adding or subtracting a proton to a neutral molecule M. Though experimentalists aim to generate only a single adduct type, in real life, multiple adducts are formed (Mahieu & Patti, 2017;

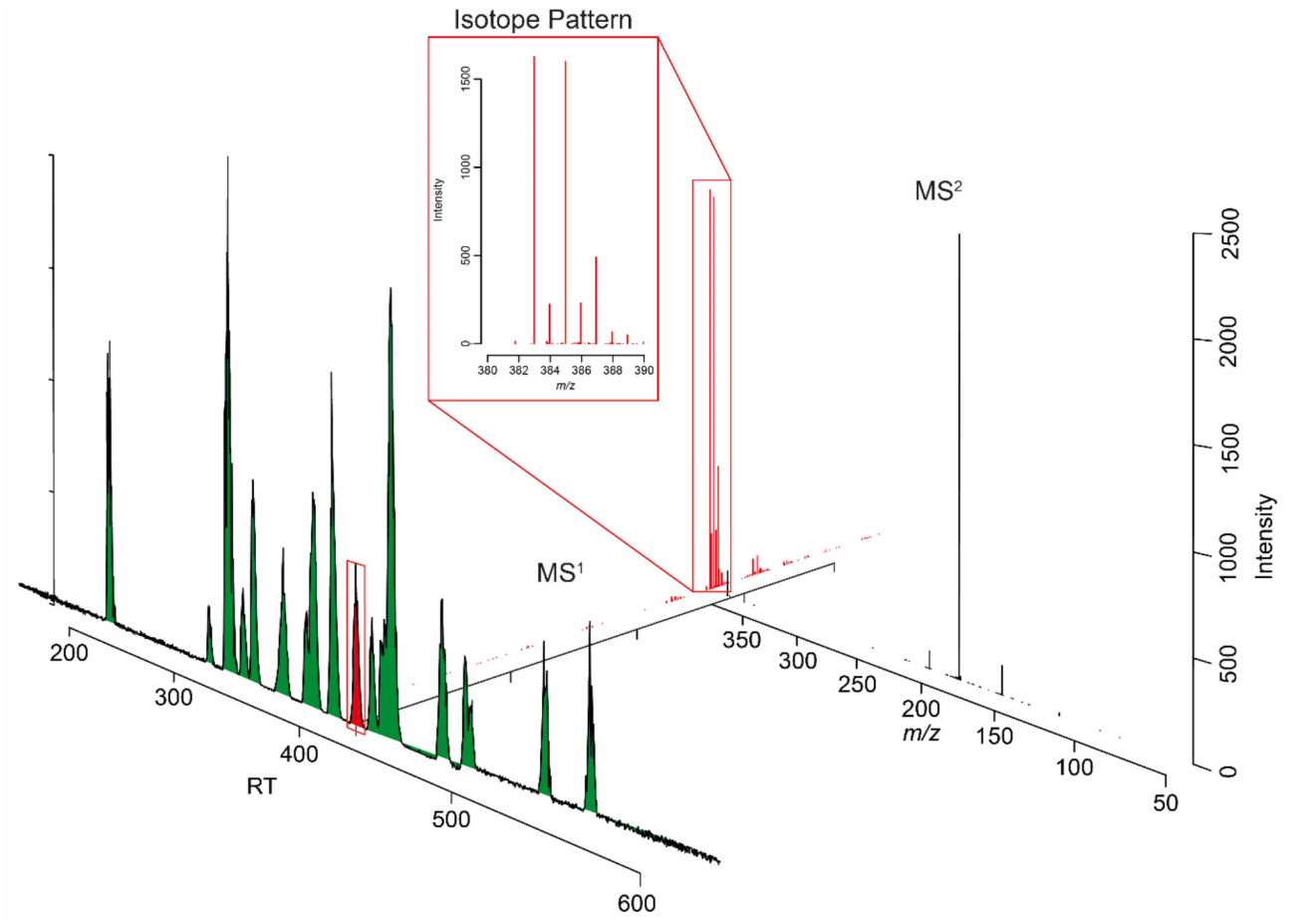

Fig. 1 Metabolomics LC-MS/MS data structure. Typical LC-MS/MS data in non-targeted metabolomics consist of a retention time and mass-to-charge dimension (MS^1) and the associated peak intensity. Additionally, one or multiple MS^2 spectra might be associated with the detected feature

Nash et al., 2024). Other examples of adducts are $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$ in positive or $[M+FA\text{-}H]^-$, $[M+HAc\text{-}H]^-$ in negative ionization mode. The extent to which the different adducts are observed depends on the structure of the molecule and the experimental conditions, such as the geometry of the ion source, ionization voltages, and the mobile phase composition. The quality of solvents is another critical factor. Water in mass spectrometry quality, stored over a longer period in glass bottles, shows a higher Na and K ions content, which can lead to higher intensities of $[M+H]^+$ and $[M+K]^+$ adducts (personal observation). On the other hand, ions like $[M+NH_4]^+$ can only be formed if NH_4^+ is present in the mobile phase. Certain metabolites require this for ionization, e.g., species from the lipid class triacylglycerols mostly ionize as $[M+NH_4]^+$. Tools are currently being developed to predict adductation and ion species but are far from being used.

In addition to adducts, in-source fragments can also be observed. These happen if conditions in the ion source or any other part of MS before the collision cell lead to fragmentation, primarily due to too high ionization voltages, temperatures, or high voltage drops in ion optics. However, this effect is also dependent on the chemical structure of the metabolite and, consequently, its stability. Therefore, in-source fragmentation is a phenomenon that is not consistent across all metabolites. An example of an in-source fragment is $[M\text{-}H_2O+H]^+$ or $[M\text{-}H_2O\text{-}H]^-$, which can be seen for molecules containing hydroxyl groups. Besides losses of small molecule parts such as hydroxyl, amino, phosphate, or sulfate groups, larger parts can also be lost in in-source fragmentation, e.g., sugar moieties. No general rules for in-source fragments can be stated since they depend on the individual metabolite structures. Sometimes, in-source fragmentation can be so extensive that no intact ion species is observed. For example, the amino acid tryptophan has a very prominent loss of the amino group, leading to an ion with the same *m/z* as indole acrylic acid (*m/z* 188.0706). Under certain conditions, only this ion and no intact $[M+H]^+$ are observed.

Additionally, for each observed ion species or adduct, isotopic peaks can be observed, depending on the intensity of the individual adduct, further complicating the spectrum.

Most elements have isotopes, which have the same number of protons but different neutrons. Metabolites are often built from carbon (C), hydrogen (H), oxygen (O), nitrogen (N), sulfur (S) and phosphor (P). In nature, the isotope ^{13}C is present at 1.1% natural abundance. Therefore, at least one isotopic peak can be observed in most cases. Other isotopes metabolomics are ^{15}N (0.4% natural abundance), ^{18}O (0.2% natural abundance), and ^{34}S (4,37% natural abundance). Beside these, isotopes of chlorine and bromine, potentially found in secondary metabolites or xenobiotics, are of high abundance (^{37}Cl (24% natural abundance) and ^{81}Br (49.4% natural abundance)).

Since isotopes behave chemically (almost) the same and adducts and in-source fragments are formed after the chromatographic separation step in the MS, they will all have the same RT, a fact that can be used for grouping signals. Furthermore, their chromatographic peak shape should be very similar (showing a high correlation across the chromatographic profile). While for isotopes and adducts, defined rules can be used for their identification (e.g., constant distance of $[M+H]^+$ and $[M+Na]^+$ adducts), this is hardly possible for in-source fragments. Data processing pipelines such as xcms, MZmine, MS-DIAL, or commercial software solutions ideally deal with different adducts, isotopes, and in-source fragments. However, especially in the case of in-source fragments, it cannot be clear if it is an in-source fragment or a molecule with a similar mass and retention time (Domingo-Almenara et al., 2019; Guo et al., 2021). To resolve this ambiguity, metabolite standards are measured to confirm ion species. For example, Fig. 2 shows the MS^1 spectrum of a tryptophan standard, illustrating different ion species that can be observed. Data were obtained from a Bruker maXis UHR-ToF-MS. For other MS, this spectrum will look different.

While different peaks in the MS^1 spectrum are informative and allow calculation of the sum formula, they do not allow for derivation of the identity of the metabolite. More detailed information is required in the form of fragmentation spectra to do so. Here, the MS fragments a pseudo-molecular ion into smaller pieces, which might be diagnostic to identify metabolites unequivocally. Differences in the exact MS setup and settings can cause major differences. To lead to fragmentation, external energy must be applied to the ions in the collision cell. This energy is typically referred to as collision energy. Besides collision energy, the actual geometry of the collision cell and the adduct type of the precursor, which shall be fragmented, influence the resulting fragmentation spectrum.

In non-targeted metabolomics, two types of instrumentation are used: Orbitraps and QTOFs. The Orbitrap instrumentation offers two types of fragmentation: CID and HCD, while QTOFs typically offer only CID fragmentation. The difference is that in Orbitraps, the CID is performed in the ion trap, while HCD is performed in the C-Trap. Notably, on Orbitraps, the CID resembles a tandem-in-time configuration type, and the HCD is a tandem-in-space configuration. Each metabolite behaves differently during fragmentation, and the optimal collision energy to obtain the most informative spectrum also changes. However, generic settings are often used since the molecule's identity is unknown in non-targeted metabolomics. Such generic settings can be suboptimal, leading to under- or over-fragmentation of metabolites that either have too little or too strong fragmentation to be informative. Therefore, multiple collision energies or stepped or ramped collision energies are often used. Figure 2C shows three different fragmentation spectra of the [M + H] + adduct of tryptophan with 10, 20, and 40 eV obtained on a Bruker maXis UHR-ToF-MS. *De novo* identification can often be improved by combining information from multiple collision energies or using ramp spectra (Hoffmann et al., 2022). Different adducts (e.g., $[M+H]^+$, $[M+Na]^+$ or $[M+NH_4]^+$) will lead to different fragmentation spectra. Besides HCD and CID for Orbitraps and CID for QTOFs as "standard" fragmentation modes, more novel fragmentation methods are on the rise for potential deeper structural insights and enhanced metabolite identification, e.g., Electron Activated Dissociation (EAD), Oxygen Attachment Dissociation (OAD) or Ultraviolet Photodissociation (UVPD).

3.1 Practical implications for the metabolomics bioinformatician

Based on the data and degree of processing performed, the feature table will vary depending on the processing stage. The features can be either ungrouped or grouped only by isotope or by isotope and adduct. It is essential to know the state of the feature table, as specific steps in data processing can be skipped or have to be performed. As a summary for a bioinformatician, each feature in the feature table (i.e., potential metabolite) will be represented as multiple rows, as when various sequences match the same gene during the alignment process of RNA-Seq analysis. Typically, isotopes are first grouped as they are the easiest to detect based on constant mass differences. Most software tools will perform this, or add-on packages exist (e.g., CAMERA for xcms (Kuhl et al., 2011)). Depending on the software or program used, the table either contains rows for isotopes, with a notation denoting them as such, or does not contain isotopic peaks. Further grouping, e.g., of adducts, depends on the software used. Though this adduct grouping is an important step, tools often perform sub-optimally using certain assumptions, e.g., the main adduct is $[M+H]^+$, and other adducts can be identified as seed. While this might hold true

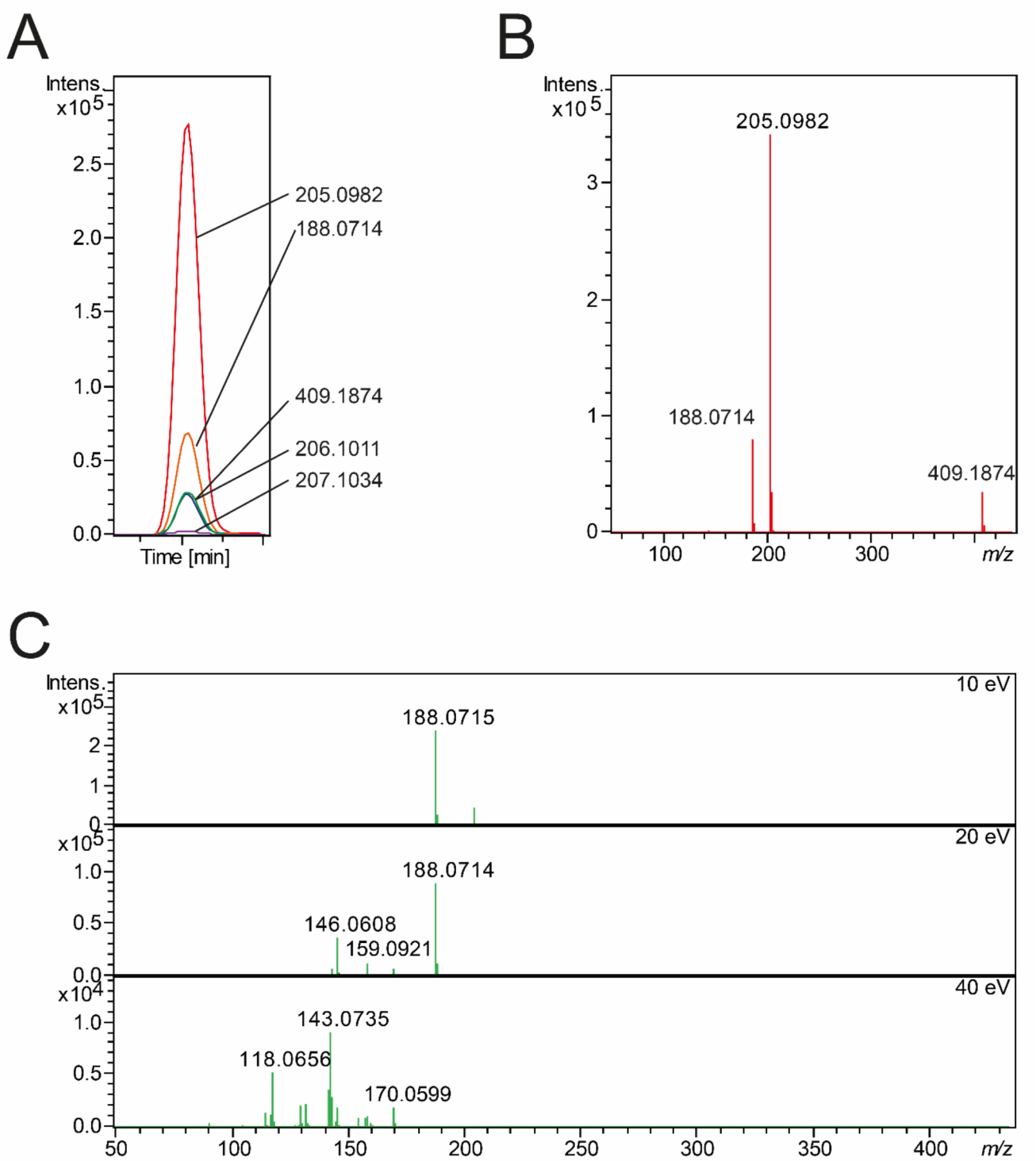

Fig. 2 (**A**) Extracted Ion Chromatograms of different m/z derived from tryptophan. m/z 205.0982, 206.1011 and 207.1034 are derived from the [M+H]+adduct and represent the monoisotopic and the first to isotopic peaks. m/z 188.0714 represents an in-source fragment and m/z 409.1874 the $[2\ M+H]^+$ adduct. (**B**) MS^1 spectrum of all peaks detected in A. (**C**) MS^2 spectra of tryptophan collected at different collision energies. All data was generated on a Bruker maXis UHR-TOF-MS

for many metabolites and grouping might work, unusual adducts or extensive in-source fragmentation might lead to incorrect grouping. In fact, correct adduct annotation can often only be established through metabolite identification (see below). Further software tools for feature table analysis might be based on similar assumptions. To become aware of these, carefully reading the documentation or vignettes associated with software tools will be necessary. Thus, the resulting feature table will need further annotation (i.e., to identify which metabolites you have) to interpret your data. However, as we will explain in the next section, annotation is quite complex.

Apart from the actual feature table, you will likely work with fragmentation spectra and further downstream analysis, such as metabolite identification. It is therefore important to familiarize yourself with software tools or packages that can handle this type of data, e.g., the R package Spectra (Rainer et al., 2022) or the Python package matchms (Huber et al., 2020).

Lastly, remember that the same metabolite measured in different machines and/or using different parameters (e.g., collision energy) will lead to different adducts, in-source fragments, and fragmentation spectra. Therefore, different datasets from different LC-MS/MS instruments will need different parameter settings for processing and feature grouping.

4 Metabolite identification is hard!

To correctly identify metabolites, a plethora of information is required, which includes MS^1 and MS^2, as well as complementary information, such as retention time (RT) and/or collisional cross sections (CCS). The major problem is that no common unifying physicochemical principle for metabolites (and lipids) can be found in DNA, RNA, or proteins. There is no sequence to which metabolite signals can be mapped. However, the measured features need to be converted to metabolites to correctly interpret non-targeted metabolomics data in a biochemical context.

Different levels of metabolite identification can be achieved (Sumner et al., 2007). One has to differentiate between identification and annotation. While the first one is used when the identity of a metabolite is certain, annotation still gives room for uncertainty. The highest level of identification is achieved by comparing signals from measured biological features against an in-house reference database obtained from chemical reference standards; therefore, identification is limited to the standards available in the laboratory. Matching should be performed based on the exact *m/z*, fragmentation pattern, and RT. For the highest accuracy, matching must be performed using the same collision energies. Lower levels of identification or annotation are achieved by comparing the feature signals against those from public databases, such as GNPS, MassBank, Metlin, etc. Here, the best results are achieved when fragmentation spectra are matched against public spectra from the same instrument and the same or similar collision energies. However, it might be unlikely that precisely the metabolites of interest are available on the employed LC-MS/MS system. Therefore, matching against any available library gives potential ideas about the metabolite identity.

Matching is often limited to MS^1 data since RTs are commonly not shared or differ too much. Different approaches have been developed to make RT sharing possible and to make it compatible between different laboratories (Aalizadeh et al., 2022; Hao et al., 2023; Renaud et al., 2021; Stoffel et al., 2022). However, coverage of publicly available spectral libraries is limited (Frainay et al., 2018). Fortunately, besides matching against various databases (in-house or public), *in silico* approaches have been developed to increase the coverage of metabolites. In such approaches, structural databases are used instead of spectral libraries. One of the first *in silico* methods developed was MetFrag, which uses combinatorial bond-breaking to find potential fragment structures matching peaks in the fragmentation spectra (Ruttkies et al., 2016). CSI: FingerID, another tool, uses machine learning techniques to learn from fragmentation trees, sum formulas, and molecular fingerprints computed from the fragmentation spectrum to annotate potential compounds (Böcker & Dührkop, 2016; Dührkop et al., 2015). Besides these two, multiple other *in silico* tools exist (Djoumbou-Feunang et al., 2019; Ridder et al., 2014). It is essential to mention that these tools are likely always to give a result, especially if large structural databases such as PubChem are used as input. Therefore, great care needs to be taken when evaluating the results of these tools. The performance of such tools is evaluated in contests like CASMI (CASMI, 2024; Kasama et al., 2014; Schymanski & Neumann, 2013; Shen et al., 2013), for example. However, high scores from *in silico* annotations do not reflect high confidence. A recent method has been established to estimate the confidence of CSI: FingerID results, and it turns out that the performance is similar to library matching (Hoffmann et al., 2022). Though this is a significant step forward, these approaches are still very limited. The use of false discovery rates (FDRs), similar to transcriptomics and proteomics, is still very limited in metabolomics.

Despite the growth of mass spectral libraries and the existence of all these *in silico* tools, annotation and identification often do not achieve the full structural detail as found in metabolite or pathway databases. This is especially true for the position and stereochemistry of double bonds in acyl chains, the position of functional groups (e.g., hydroxyl

groups), or stereochemistry in general. For instance, typical RP- or HILIC-MS/MS methods which are employed in non-targeted metabolomics, cannot differentiate between D- or L-Tryptophan, but would require dedicated chiral separation methods (Müller et al., 2014). Fragmentation spectra of both stereoisomers are identical, and retention time cannot be differentiated unless specific chiral chromatography is used. In such cases, one form is often assumed (e.g., the L-form) to be the canonical version in mammals. However, this usually does not hold true when working with bacteria or gut microbiome samples. In the future, pathway analysis tools need to cope with this uncertainty.

4.1 Practical implications for the metabolomics bioinformatician

Best identifications and annotations are obtained using reference libraries measured on the same instrument under the same conditions as the samples of interest. However, the availability of reference standards is limited, and a laboratory is unlikely to hold standards for all metabolites of interest. Significant efforts have been made to increase the number of structures covered in public databases, but directly using a public database would be too easy. In reality, although fragmentation spectra of the same metabolite measured in different machines and/or with different parameters (e.g., collision energy) will generally exhibit an overall similar pattern, they will not be identical. This means that, for the identification to be reliable, the standards should be measured in the same machine (not only one of the same brand and model but literally the same machine) and with the same parameters used to process the samples you are analyzing.

Depending on the organism you are studying, the corresponding metabolites will be known (as when you have a genome of reference in transcriptomics) or not (as in de-novo transcriptome assembly). Even if a metabolite is known, it must have corresponding fragmentation spectra deposited. In the case of databases for metabolite identification and annotation, one needs to differentiate between mass spectral databases holding information on metabolite structures and corresponding tandem MS spectra and metabolite or structural databases containing only structure information. Examples of MS databases are MassBank or GNPS, while KEGG, HMDB, and ChEBI are primary structural databases (though HMDB also contains many spectra from MS and NMR).

Different *in silico* tools such as MetFrag or CSI: FingerID allow the search with MS/MS data in structural databases. Although these tools are improving, their results represent only annotations and not identifications. They need to be treated with care and have to be confirmed by an analytical chemist, ideally using a chemical reference standard. But even with all the tools and databases available, don't get your hopes up and think that you will annotate every single feature or even most of them, as in RNA-Seq; in most cases, annotation rates in metabolomics are way below 10%, and the remaining 90% are often referred to as the "dark matter" in metabolomics (da Silva et al., 2015).

5 Metabolite naming and identifiers are not always unique

The chemical structure of a metabolite is its most unique identifier. Different ways to store the structure electronically exist, e.g., .sdf (structures data file). However, this format is hard to use when working with tabular data. String representations of structures are used in bio- and cheminformatics, e.g., the Simplified Molecular Input Line Entry System (SMILES) or IUPAC International Chemical Identifier (InChI). A hashed version of the InChI, the InChIKey, exists, which is a fixed-length digital representation. Theoretically, the InChIKey is collision-free and, therefore, can serve as a unique identifier. InChIs and InChIKeys can only be generated for molecules with completely known atom connectivity, while SMILES allows for some degree of uncertainty (e.g., exact acyl side chain in a lipid structure). To work with metabolomics and lipidomics results, it is essential to familiarize yourself with the concepts behind SMILES, InChI, and InChIKey, as they are often used in reports. InChIKey can be used to map chemical structures in different datasets, e.g., using only the first of three layers (atom connectivity) with certain restrictions (e.g., all hexoses, such as glucose, fructose, etc…, will have the same first layer). To avoid mismatching due to different representation, it is advised to normalize chemical structures before generating InChIs and InChIKeys, e.g., using the PubChem normalization (Hähnke et al., 2018).

From the structure, the name and chemical formula can be derived. Different names for the same chemicals exist. For example, the chemical IUPAC name of L-Tryptophan is (2 S)-2-amino-3-(1 H-indol-3-yl)propanoic acid. PubChem lists close to 250 synonyms for this single metabolite. Different metabolite databases exist, with the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Sakurai et al.), Chemical Entities of Biological Interest (ChEBI) (Hastings et al., 2016), Human Metabolome Database (HMDB) (Wishart et al., 2021) and Lipid Maps (Liebisch et al., 2020) representing the most used ones. Apart from these, specialized databases that are specific to some organisms exist. For example, SMID-DB stores information on secondary metabolites from the nematode *Caenorhabditis elegans* (Artyukhin et al., 2018). Furthermore, genome-scale

metabolic models represent knowledge bases for the metabolism of a given organism. Different tools for analyzing metabolomics results might use different metabolite databases and, therefore, require specific database identifiers. No unified metabolite database as a de-facto standard (similar to UniProt for proteins, for example) currently exists.

Furthermore, the structural details of the identified metabolites might differ from those in the respective databases. For example, while based on standard metabolomics approaches, only tryptophan can be identified (no stereochemistry, see above), the database may contain L- and D-Tryptophan. Additionally, in genome-scale metabolic networks, metabolites might be present at a different charge state, as these models are typically mass- and charge-balanced for microspecies at pH 7.3. ChEBI, for example, stores different versions of metabolites, such as with or without defined stereochemistry or different charge states as separate entities. These individual entities are linked to each other by a rich ontology, and entries in different databases might be connected to each other via cross-references. For example, the KEGG compound C16434, called isoleucine, is linked to ChEBI:38,264, called 2-amino-3-methylpentanoic acid. Though this is correct, as isoleucine is a 2-amino-3-methylpentanoic acid, there also exists a ChEBI:24,898 called isoleucine, which would be the correct link.

5.1 Practical implications for the metabolomics bioinformatician

Individual metabolites might be named by different chemical names or identifiers, which often refer to the same or very similar structure. As a bioinformatician, be aware that not every metabolite database will contain all metabolites; IDs will differ, and direct mapping to databases and metabolic pathways might not be possible due to discrepancies between the level of detail of the identified metabolite and the metabolite in the database/pathway. Different tools for the conversion of identifiers exist and can be incorporated into data analysis pipelines. Depending on the downstream tool used to interpret the results, different identifiers might be required. Therefore, mapping between different databases will be required depending on the results obtained from metabolite identification. Due to the large number of possible names, it is very hard to create automatic workflows to map chemical names. It is advisable to use dictionaries, established databases, and mapping tools. Such mapping can be done with different tools, such as the Chemical Translation Service (CTS) (Wohlgemuth et al., 2010), RefMet (Fahy & Subramaniam, 2020), BridgeDB (van Iersel et al., 2010), and UniChem (Chambers et al., 2013). However, not all metabolites might be present in all databases, which has particular implications, for example, for overrepresentation analysis (see below). In order to generate reproducible results, it is important to work with accurate lists of metabolites present in organisms (e.g., from genome-scale metabolic models) and prepare tables with identifiers, names, etc., in advance. Likewise, similar tables shall be provided by the analytical scientists for the metabolites identified to avoid tedious formatting and conversion. In case of ambiguities, a biochemist and/or an analytical chemist can help identify the most probable match.

Mapping between databases or datasets can be performed using the InChIKey or at least the first block (atom connectivity) to account for potential differences in stereochemistry and charge states. Furthermore, the ontology established by ChEBI can be used for mapping (Poupin et al., 2020). This mapping approach can account for certain ambiguities and was used for example, to map lipids to genome-scale metabolic models. A distance measure is reported based on the distance in the ChEBI ontology. However, this approach is only suitable for entries well established in the ontology, and mapping distances can only be partially used as a quantitative metric.

6 Metabolite coverage is low

Typical non-targeted metabolomics can detect several hundred to thousands of features, but only a few hundred metabolites or fewer can be annotated or identified, and even fewer can be detected at the highest confidence level (confirmed by a chemical reference standard). Metabolites span an extensive range of polarity and concentrations. Currently, no single analytical method can cover the entire metabolome. For example, reversed-phase LC-MS/MS (RP-MS/MS) can cover non-polar metabolites such as fatty acids, acyl-carnitines, or bile acids, or even complex lipids, Hydrophilic Liquid Interaction Chromatography (HILIC-MS/MS covers polar metabolites such as amino acids, amines, sugars, and others. Therefore, depending on the employed methodology, only a small subfraction might be sampled and analyzed. To achieve greater coverage, multiple methods need to be combined.

In many cases, positive and negative ionization modes must be combined to detect sufficient metabolites. For example, single reactions can change the properties quite substantially. For example, the metabolites glutamine, glutamic acid, and 2-oxo-glutaric acid are connected by a linear chain of reactions that feeds into the TCA cycle. However, their physicochemical properties differ. While all of them can be detected in negative ionization mode, only glutamic acid and glutamine are typically detected in positive ionization mode (see Fig. 3A).

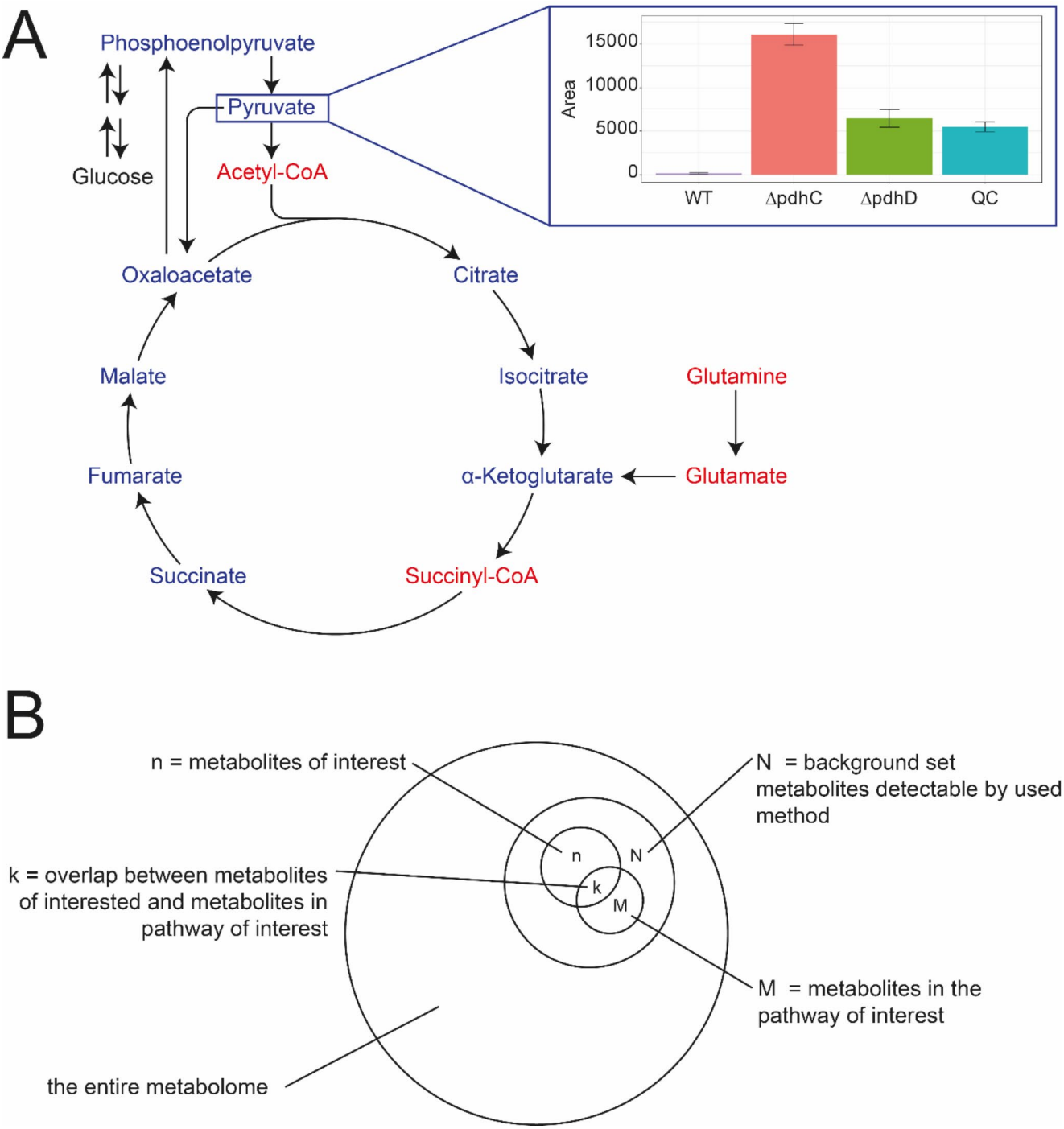

Fig. 3 (**A**) Detection of metabolite depends on their physicochemical properties. Certain metabolites can be only detected in negative mode (blue), while other are preferentially detected in positive ionization mode. However, even if a metabolite is theoretically detectable, it might be not present in samples. The example shows relative levels of pyruvic acid in Bacillus subtilis and can be only detected if the downstream consuming enzyme is mutated. (**B**) Metabolite coverage is important for analysis methods such as overrepresentation analysis. The entire metabolome cannot be covered and typically only a subset N is detectable by a specific method. This subset represents the background set for overrepresentation analysis. Figure adopted from Wiederer et al

Additionally, the metabolome is highly dynamic. Even if metabolites are known and in theory, can be detected by the employed methods (e.g., by measuring a chemical reference standard), they might not be detected because they are either not produced under the given biological conditions or metabolized too fast. Specific metabolites have very high turnover rates and might not be detected when the overall turnover in the reaction cascade is very high. However, the

exact flux and turnover are dependent on many different factors, such as growth conditions and nutrient availability. For example, pyruvic acid, an important intermediate between glycolysis and the TCA cycle, can only be detected in *Bacillus subtilis* if the downstream-consuming enzyme (pyruvate dehydrogenase) is mutated, since the flux into further downstream pathways is disturbed and the metabolite accumulates (see Fig. 3A). Furthermore, specific metabolites might not be stable enough to be detected or yield reliable readouts, including central metabolites such as ATP, NADH or NADPH. Similarly, some metabolites may be present at very low concentrations. The metabolome can span several orders of magnitude in concentration (Bennett et al., 2009). For very low-concentration metabolites, specialized assays with dedicated sample preparation (e.g., solid-phase extraction) and targeted analysis using triple quadrupole MS are often required, rendering them undetectable in non-targeted metabolomics settings. Lastly, for almost all organisms, the complete metabolome remains unknown, so the actual coverage of a method cannot be precisely calculated, which also has implications for the downstream interpretation of results.

6.1 Practical implications for the metabolomics bioinformatician

While transcriptomics or proteomics often produce a single data table for analysis with sufficient coverage of several hundred to thousands of transcripts or proteins, metabolomics often yields multiple feature tables. Depending on the study, these feature tables might need to be joined or analyzed separately. One example is the same chromatographic method's positive and negative ionization mode. Running multiple methods may be necessary to achieve sufficient coverage of metabolites. Often, this approach is also the case in targeted metabolomics. For example, even though only a single data table is reported, the actual measurement for the commercial Biocrates MxP Quant 500 kit utilizes four different methods. The quality and coverage of metabolomics measurements significantly influence the downstream data analysis and interpretation.

Pathway analysis is one of the major data analysis techniques for the interpretation of metabolomics data, mainly in the form of enrichment or overrepresentation analysis. In contrast to such analyses used in transcriptomics and proteomics, pathway analysis in metabolomics requires great care. The detection of metabolites can be highly biased by the employed analytical method. Therefore, selecting the correct background dataset is important. While the number of detected features is typically very large for transcriptomics and proteomics, and the entire genome or proteome can be used as a background set, this often is not the case in metabolomics. Recently, Wieder et al. published recommendations for the overrepresentation analysis in metabolomics (Wieder et al., 2021). These recommendations included selecting the background dataset, pathway database, and addressing potential misidentifications. To overcome this limitation in coverage, methods for integrating multiple datasets from different methods will be required. While this integration can be achieved for targeted metabolomics and known metabolites, e.g., by metrics established by Boccard et al., similar integrations of multiple datasets of unknown metabolites remain complicated (Boccard & Rudaz, 2018).

Since the coverage of metabolic pathways is often limited by the number of identified metabolites, alternative analysis methods are required. Network analysis has emerged as a suitable alternative to overcome the limitations of pathway analysis. Different types of networks provide different views of biological aspects and can even be integrated with each other (Amara et al., 2022; Salzer et al., 2023).

7 Conclusion

Compared to genomics, transcriptomics, and proteomics, metabolomics, and lipidomics are missing one common part: sequences that can be matched against each other. These sequences make it possible to map reads in RNA-seq to the respective genes or lead to specific fragmentation of peptide backbones in peptides, allowing for their sequencing. Furthermore, false discovery rates can be calculated for them, allowing one to judge the goodness of the sequencing result and mapping. Metabolites have no common characteristics that can be used in a similar fashion, making metabolite analysis seemingly complicated from a bioinformatic point of view.

There is a strong need for improved data analytical and bioinformatic tools in metabolomics. These include steps such as raw data processing, metabolite identification, and data interpretation. Though we highlighted several pitfalls and problems in this review, there is great progress in the development of new tools.

In our opinion, teaching in bioinformatics is mostly sequence-based and covers genomics, transcriptomics, and proteomics. Small molecules such as metabolites and lipids, are often not covered or are covered only superficially. Since this is an upcoming topic and more and more metabolomics data will be generated, structured training is required. In this article, only the view of the bioinformatician on specific aspects are covered. However, the other side also needs to adapt. Training for analytical chemists or biochemists needs to involve more basic data science and bioinformatics skills. Data generated by the latest state-of-the-art instrumentation is very dense and requires advanced skills. Despite major

challenges still existing, there are many useful resources, protocol papers and guides (Amara et al., 2022; Aron et al., 2020; Heuckeroth et al., 2024; Pakkir et al., 2023). It is important to note that many of these articles and tutorials have been written by analytical chemists and bioinformaticians in collaboration with exactly the goal of bringing together the two fields. Though it seems from this review that often there are more problems than solutions, the authors always found a way to improve their collaboration and overcome all hurdles.

Here are some necessary steps as a summary for (future) metabolomics bioinformaticians to avoid the same pitfalls as the authors:

1. Team up with the (bio)chemist who generated the data as early as possible (best before the data has been generated); it will make your life much easier. Ask if you can join/watch the measurements. Ask many, many questions about the data.
2. Start by figuring out the stage the feature table is in because the steps to follow will depend on this. Examples of useful questions to ask are "Is the feature table grouped or ungrouped?" If it is grouped, "What was used for the grouping: only isotopes or isotopes and adducts?".
3. Remember that it is common to have multiple rows representing the same feature, and even if your feature table is grouped, the grouping might be incorrect (e.g., due to extensive in-source fragmentation). Additionally, when working with multiple datasets, keep in mind that measurements performed in different machines and/or using different sets of parameters lead to different features. This is relevant for processing and feature grouping (your parameter settings should be adapted for each dataset) and for the annotation/identification step.
4. When annotating your features, keep in mind the difference between identification (knowing the metabolite with certainty) and annotation (finding the metabolite that is likely to represent the feature). You might as well check out some MS databases, such as MassBank and GNPS, as well as tools like MetFrag or CSI: FingerID for *in silico* annotation. Please remember that ~90% of the features are expected to remain "unknown"; i.e., you'll be able to annotate less than 10% of the features.
5. Be aware that the metabolite databases are likely to be incomplete. To complicate things further, the same metabolite can have different names and/or identifiers depending on the database, which does not exactly help when mapping between databases. We, therefore, recommend using predefined dictionaries, established databases, and mapping tools (such as CTS, RefMet, BridgeDB, and UniChem). Candidates to perform such mapping are the InChiKey and the ChEBI.
6. Familiarize yourself with the type of data you'll need to process (for instance, the feature table and the fragmentation spectra of some or all of the features) and the tools you will need to process it, for example, R packages such as xcms and Spectra are very likely to be useful.
7. By this point, you should already be used to dealing with challenges and uncertainty. Things won't get any easier for the interpretation step (e.g., via enrichment or overrepresentation analysis). Still, you can take into account the recommendations from Wieder et al. [35] and the metrics established by Bocard et al. [36]. Finally, you can also consider a network approach to interpret your results. As we explored in the MetClassNet project, different networks can be integrated to provide various biological perspectives on the same dataset.

Last but not least, metabolomics data analysis is fun and can be very rewarding once you accept the points above. Increasing the coverage by combining methods, creating better and faster methods, and improving metabolite identification rates are current fields of research in the metabolomics community and are expected to improve significantly in the coming years. There are still many pitfalls, but so far it has been successful, and the field is growing with new solutions and tools being released and published every year.

Acknowledgements A huge thank you to all members of the MetClassNet project and our collaboration partners outside of this project. It has been a lot of fun to work with you all. "Merci beaucoup" and "Vielen Dank".

Author contributions Both authors have jointly written this review. Figures have been created by Michael Witting.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the Agence Nationale de la Recherche (ANR, French National Research Agency)—MetaboHUB, the national metabolomics and fluxomics infrastructure (Grant ANR-INBS-0010), Project number ANR-19-CE45-0021 (MetClassNet) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project number 431572533 (MetClassNet).
Open Access funding enabled and organized by Projekt DEAL.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Aalizadeh, R., Nikolopoulou, V., & Thomaidis, N. S. (2022). Development of Liquid Chromatographic Retention Index based on Cocamide Diethanolamine homologous series (C(n)-DEA). *Analytical Chemistry*, *94*, 15987–15996.

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., Ewald, C., Fraser, J., Giavalisco, P. D., Hall, P., Heinemann, R. D., Link, M., Luo, H., Neumann, J., Nielsen, S., Perez, J., de Souza, L., Saito, K., Sauer, U., Schroeder, F. C., Schuster, S., Siuzdak, G., Skirycz, A., Sumner, L. W., Snyder, M. P., Tang, H., Tohge, T., Wang, Y., Wen, W., Wu, S., Xu, G., Zamboni, N., & Fernie, A. R. (2021). Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices. *Nature Methods*, *18*, 747–756.

Amara, A., Frainay, C., Jourdan, F., Naake, T., Neumann, S., Novoa-del-Toro, E. M., Salek, R. M., Salzer, L., Scharfenberg, S., & Witting, M. (2022). Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Frontiers in Molecular Biosciences* 9.

Aron, A. T., Gentry, E. C., McPhail, K. L., Nothias, L. F., Nothias-Esposito, M., Bouslimani, A., Petras, D., Gauglitz, J. M., Sikora, N., Vargas, F., van der Hooft, J. J. J., Ernst, M., Kang, K. B., Aceves, C. M., Caraballo-Rodríguez, A. M., Koester, I., Weldon, K. C., Bertrand, S., Roullier, C., Sun, K., Tehan, R. M., Boya, P., Christian, C. A., Gutiérrez, M. H., Ulloa, M., Mora, A. M. T., Mojica-Flores, J. A., Lakey-Beitia, R., Vásquez-Chaves, J., Zhang, V., Calderón, Y., Tayler, A. I., Keyzers, N., Tugizimana, R. A., Ndlovu, F., Aksenov, N., Jarmusch, A. A., Schmid, A. K., Truman, R., Bandeira, A. W., Wang, N., M. and, & Dorrestein, P. C. (2020). Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols*, *15*, 1954–1991.

Artyukhin, A. B., Zhang, Y. K., Akagi, A. E., Panda, O., Sternberg, P. W., & Schroeder, F. C. (2018). Metabolomic Dark Matter Dependent on Peroxisomal β-Oxidation in Caenorhabditis elegans. *Journal of the American Chemical Society*, *140*, 2841–2852.

Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., & Rabinowitz, J. D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nature Chemical Biology*, *5*, 593–599.

Benton, H. P., Wong, D. M., Trauger, S. A., & Siuzdak, G. (2008). XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization. *Analytical Chemistry*, *80*, 6382–6389.

Boccard, J., & Rudaz, S. (2018). Chapter Seventeen - Integration of Metabolomic Data From Multiple Analytical Platforms: Towards Extensive Coverage of the Metabolome in Jaumot, J., Bedia, C. and Tauler, R. (Eds.), *Comprehensive Analytical Chemistry*, Elsevier. pp. 477–504.

Böcker, S. (2019). Algorithimic Mass Spectrometry: From molecules to masses and back again.

Böcker, S., & Dührkop, K. (2016). Fragmentation trees reloaded. *Journal of Cheminformatics*, *8*, 5.

CASMI (2024). CASMI 2022 Results.

Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S., & Overington, J. P. (2013). UniChem: A unified chemical structure cross-referencing and identifier tracking system. *Journal of Cheminformatics*, *5*, 3.

da Silva, R. R., Dorrestein, P. C., & Quinn, R. A. (2015). Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, *112*, 12549–12550.

Djoumbou-Feunang, Y., Pon, A., Karu, N., Zheng, J., Li, C., Arndt, D., Gautam, M., Allen, F., & Wishart, D. S. (2019). CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification. *Metabolites*, *9*, 72.

Domingo-Almenara, X., Montenegro-Burke, J. R., Guijas, C., Majumder, E. L. W., Benton, H. P., & Siuzdak, G. (2019). Autonomous METLIN-Guided In-source fragment annotation for untargeted metabolomics. *Analytical Chemistry*, *91*, 3246–3253.

Dührkop, K., Shen, H., Meusel, M., Rousu, J., & Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of National Academy of Sciences* 112.

Fahy, E., & Subramaniam, S. (2020). RefMet: A reference nomenclature for metabolomics. *Nature Methods*, *17*, 1173–1174.

Frainay, C., Schymanski, E. L., Neumann, S., Merlet, B., Salek, R. M., Jourdan, F., & Yanes, O. (2018). Mind the gap: Mapping Mass Spectral databases in Genome-Scale metabolic networks reveals poorly covered areas. *Metabolites*, *8*, 51.

Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G. G., Salek, R. M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q. W., del Toro, N., Pérez-Riverol, Y., Ghali, F., Bandeira, N., Xenarios, I., Kohlbacher, O., Vizcaíno, J. A., & Hermjakob, H. (2014). The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental results to a wider audience *. *Molecular & Cellular Proteomics*, *13*, 2765–2775.

Guo, J., Shen, S., Xing, S., Yu, H., & Huan, T. (2021). ISFrag: De Novo Recognition of In-Source fragments for Liquid Chromatography–Mass Spectrometry Data. *Analytical Chemistry*, *93*, 10243–10250.

Hähnke, V. D., Kim, S., & Bolton, E. E. (2018). PubChem chemical structure standardization. *Journal of Cheminformatics*, *10*, 36.

Hao, J. D., Chen, Y. Y., Wang, Y. Z., An, N., Bai, P. R., Zhu, Q. F., & Feng, Y. Q. (2023). Novel peak shift correction method based on the Retention Index for Peak Alignment in Untargeted Metabolomics. *Analytical Chemistry*, *95*, 13330–13337.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, *44*, D1214–D1219.

Heuckeroth, S., Damiani, T., Smirnov, A., Mokshyna, O., Brungs, C., Korf, A., Smith, J. D., Stincone, P., Dreolin, N., Nothias, L. F., Hyötyläinen, T., Orešič, M., Karst, U., Dorrestein, P. C., Petras, D., Du, X., van der Hooft, J. J. J., Schmid, R., & Pluskal, T. (2024). Reproducible mass spectrometry data processing and compound annotation in MZmine 3. *Nature Protocols*.

Hoffmann, N., Rein, J., Sachsenberg, T., Hartler, J., Haug, K., Mayer, G., Alka, O., Dayalan, S., Pearce, J. T. M., Rocca-Serra, P., Qi, D., Eisenacher, M., Perez-Riverol, Y., Vizcaíno, J. A., Salek, R. M., Neumann, S., & Jones, A. R. (2019). mzTab-M: A Data Standard for sharing quantitative results in Mass Spectrometry Metabolomics. *Analytical Chemistry*, *91*, 3302–3310.

Hoffmann, M. A., Nothias, L. F., Ludwig, M., Fleischauer, M., Gentry, E. C., Witting, M., Dorrestein, P. C., Dührkop, K., & Böcker, S. (2022). High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology*, *40*, 411–421.

Huber, F., Verhoeven, S., Meijer, C., Spreeuw, H., Castilla, E. M. V., Geng, C., Hooft, J. J., Rogers, S., Belloum, A., Diblen, F., &

Spaaks, J. H. (2020). Matchms - processing and similarity evaluation of mass spectrometry Da Ta. *Journal of Open Source Software*, *5*, 2411.

Johnson, C. H., Ivanisevic, J., Benton, H. P., & Siuzdak, G. (2015). Bioinformatics: The Next Frontier of Metabolomics. *Analytical Chemistry*, *87*, 147–156.

Kasama, T., Kinumi, T., Makabe, H., Matsuda, F., Miura, D., Miyashita, M., Nakamura, T., Tanaka, K., Yamamoto, A., & Nishioka, T. (2014). Winners of CASMI2013: automated tools and challenge data. *Mass Spectrom* 3.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., & Neumann, S. (2011). CAMERA: An Integrated Strategy for Compound Spectra Extraction and annotation of Liquid Chromatography/Mass Spectrometry Data sets. *Analytical Chemistry*, *84*, 283–289.

Liebisch, G., Fahy, E., Aoki, J., Dennis, E. A., Durand, T., Ejsing, C. S., Fedorova, M., Feussner, I., Griffiths, W. J., Köfeler, H., Merrill, A. H. Jr., Murphy, R. C., O'Donnell, V. B., Oskolkova, O., Subramaniam, S., Wakelam, M. J. O., & Spener, F. (2020). Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures. *Journal of Lipid Research*, *61*, 1539–1555.

Mahieu, N. G., & Patti, G. J. (2017). Systems-Level annotation of a Metabolomics Data Set reduces 25 000 features to fewer than 1000 unique metabolites. *Analytical Chemistry*, *89*, 10397–10406.

Misra, B. B. (2021). New software tools, databases, and resources in metabolomics: Updates from 2020. *Metabolomics*, *17*, 49.

Müller, C., Fonseca, J. R., Rock, T. M., Krauss-Etschmann, S., & Schmitt-Kopplin, P. (2014). Enantioseparation and selective detection of D-amino acids by ultra-high-performance liquid chromatography/mass spectrometry in analysis of complex biological samples. *Journal of Chromatography A*, *1324*, 109–114.

Nash, W. J., Ngere, J. B., Najdekr, L., & Dunn, W. B. (2024). Characterization of Electrospray Ionization Complexity in Untargeted Metabolomic Studies. *Analytical Chemistry*.

Pakkir Shah A. K., Walter, A., Ottosson, F., Russo, F., Navarro-Díaz, M., & Boldt, J. (2023) The Hitchhiker's guide to statistical analysis of feature-based molecular networks from non-targeted metabolomics data. *ChemRxiv*. https://doi.org/10.26434/chemrxiv-2023-wwbt0

Poupin, N., Vinson, F., Moreau, A., Batut, A., Chazalviel, M., Colsch, B., Fouillen, L., Guez, S., Khoury, S., Dalloux-Chioccioli, J., Tournadre, A., Le Faouder, P., Pouyet, C., Van Delft, P., Viars, F., Bertrand-Michel, J., & Jourdan, F. (2020). Improving lipid mapping in genome scale metabolic networks using ontologies. *Metabolomics*, *16*, 44.

Rainer, J., Vicini, A., Salzer, L., Stanstrup, J., Badia, J. M., Neumann, S., Stravs, M. A., Hernandes, V., Gatto, V., Gibb, L., S. and, & Witting, M. (2022). A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. *Metabolites* 12, 173.

Renaud, J. B., Hoogstra, S., Quilliam, M. A., & Sumarah, M. W. (2021). Normalization of LC-MS mycotoxin determination using the N-alkylpyridinium-3-sulfonates (NAPS) retention index system. *Journal of Chromatography A*, *1639*, 461901.

Ridder, L., Hooft, J. J. J., & Verhoeven, S. (2014). Automatic compound annotation from mass spectrometry data using MAGMa. *Mass Spectrom* 3.

Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., & Neumann, S. (2016). MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, *8*, 3.

Sakurai, N., Ara, T., Ogata, Y., Sano, R., Ohno, T., Sugiyama, K., Hiruta, A., Yamazaki, K., Yano, K., Aoki, K., Aharoni, A., Hamada, K., Yokoyama, K., Kawamura, S., Otsuka, H., Tokimatsu, T., Kanehisa, M., Suzuki, H., & Saito, K. and Shibata, D. KaPPA-View4: A metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. *Nucleic Acids Research* 39, D677–D684.

Salzer, L., Novoa-del-Toro, E. M., Frainay, C., Kissoyan, K. A. B., Jourdan, F., Dierking, K., & Witting, M. (2023). APEX: an Annotation Propagation Workflow through Multiple Experimental Networks to Improve the Annotation of New Metabolite Classes in Caenorhabditis elegans. *Analytical Chemistry*.

Schmid, R., Heuckeroth, S., Korf, A., Smirnov, A., Myers, O., Dyrlund, T. S., Bushuiev, R., Murray, K. J., Hoffmann, N., Lu, M., Sarvepalli, A., Zhang, Z., Fleischauer, M., Dührkop, K., Wesner, M., Hoogstra, S. J., Rudt, E., Mokshyna, O., Brungs, C., Ponomarov, K., Mutabdžija, L., Damiani, T., Pudney, C. J., Earll, M., Helmer, P. O., Fallon, T. R., Schulze, T., Rivas-Ubach, A., Bilbao, A., Richter, H., Nothias, L. F., Wang, M., Orešič, M., Weng, J. K., Böcker, S., Jeibmann, A., Hayen, H., Karst, U., Dorrestein, P. C., Petras, D., Du, X., & Pluskal, T. (2023). Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology*, *41*, 447–449.

Schymanski, E., & Neumann, S. (2013). CASMI: and the winner is… *Metabolites* 3.

Shen, H., Zamboni, N., Heinonen, M., & Rousu, J. (2013). Metabolite identification through machine learning—tackling CASMI challenge using FingerID. *Metabolites* 3.

Stoffel, R., Quilliam, M. A., Hardt, N., Fridstrom, A., & Witting, M. (2022). N-Alkylpyridinium sulfonates for retention time indexing in reversed-phase-liquid chromatography-mass spectrometry–based metabolomics. *Analytical and Bioanalytical Chemistry*, *414*, 7387–7398.

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

Tsugawa, H., Ikeda, K., Takahashi, M., Satoh, A., Mori, Y., Uchino, H., Okahashi, N., Yamada, Y., Tada, I., Bonini, P., Higashi, Y., Okazaki, Y., Zhou, Z., Zhu, Z. J., Koelmel, J., Cajka, T., Fiehn, O., Saito, K., Arita, M., & Arita, M. (2020). A lipidome atlas in MS-DIAL 4. *Nature Biotechnology*, *38*, 1159–1163.

van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B. R., & Evelo, C. T. (2010). The BridgeDb framework: Standardized access to gene, protein and metabolite identifier mapping services. *Bmc Bioinformatics*, *11*, 5.

Wieder, C., Frainay, C., Poupin, N., Rodríguez-Mier, P., Vinson, F., Cooke, J., Lai, R. P. J., Bundy, J. G., Jourdan, F., & Ebbels, T. (2021). Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. *PLOS Computational Biology*, *17*, e1009105.

WishartD.S., GuoA., OlerE., WangF., AnjumA., PetersH., DizonR., SayeedaZ., TianS., LeeBrian L., BerjanskiiM., MahR., YamamotoM., JovelJ., Torres-CalzadaC., Hiebert-GiesbrechtM., LuiVicki W., VarshaviD., VarshaviD., AllenD., ArndtD., KhetarpalN., SivakumaranA., HarfordK., SanfordS., YeeK., CaoX., BudinskiZ., LiigandJ., ZhangL., ZhengJ., MandalR., KaruN., DambrovaM., SchiöthHelgi B., GreinerR., & GautamV. (2021). HMDB 5.0: The human metabolome database for 2022. *Nucleic Acids Research*, *50*, D622–D631.

Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T., & Fiehn, O. (2010). The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, *26*, 2647–2648.

Zhou, B., Xiao, J. F., Tuli, L., & Ressom, H. W. (2012). LC-MS-based metabolomics. *Molecular Biosystems*, *8*, 470–481.