ANNUAL REVIEWS

Annual Review of Analytical Chemistry

Current Challenges and Recent Developments in Mass Spectrometry–Based Metabolomics

Stephanie L. Collins,[1] Imhoi Koo,[2,3] Jeffrey M. Peters,[2] Philip B. Smith,[3] and Andrew D. Patterson[2]

[1]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

[2]Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; email: adp117@psu.edu

[3]The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Annu. Rev. Anal. Chem. 2021. 14:467–87

The *Annual Review of Analytical Chemistry* is online at anchem.annualreviews.org

https://doi.org/10.1146/annurev-anchem-091620-015205

Keywords

mass spectrometry, multi-stage mass spectrometry, metabolomics, chromatography, isotope tracing, retention indices

Abstract

High-resolution mass spectrometry (MS) has advanced the study of metabolism in living systems by allowing many metabolites to be measured in a single experiment. Although improvements in mass detector sensitivity have facilitated the detection of greater numbers of analytes, compound identification strategies, feature reduction software, and data sharing have not kept up with the influx of MS data. Here, we discuss the ongoing challenges with MS-based metabolomics, including de novo metabolite identification from mass spectra, differentiation of metabolites from environmental contamination, chromatographic separation of isomers, and incomplete MS databases. Because of their popularity and sensitive detection of small molecules, this review focuses on the challenges of liquid chromatography-mass spectrometry–based methods. We then highlight important instrumentational, experimental, and computational tools that have been created to address these challenges and how they have enabled the advancement of metabolomics research.

1. INTRODUCTION

Metabolomics, the study of the small-molecule metabolites in a system, is a widely used and powerful approach to understand the metabolic activity of organisms. It has been applied to understand the impact of human diseases and genetic perturbations on metabolism (1, 2), identify biomarkers for drug and toxicant exposure or disease (3, 4), establish the mechanism of communication between commensal or competing organisms (5), and more (6–8). The advantage of metabolomics is its ability to capture the complexity and interconnectedness of metabolic pathways in a living system, rather than study individual metabolites. Often, the source of unexpected effects of a dietary change or pharmaceutical drug on an organism can be revealed through metabolomics that would otherwise be unexplained in a reductionist approach. For example, metabolomics has revealed that antibiotics often have secondary mechanisms of microbial toxicity beyond their role in perturbing protein or nucleotide synthesis (9). Therefore, there is tremendous value in using a systems-wide approach to generate new hypotheses and capture the diversity of metabolism.

Compared to other omics techniques, metabolomics is uniquely challenging. The chemical complexity of known metabolites existing in nature is immense and continues to grow as new environments and organisms are studied. According to the Kyoto Encyclopedia of Genes and Genomes (KEGG), there are over 16,000 known unique enzymes (10) that can each produce many metabolites, depending on the substrates they accommodate. The metabolome does not consist of a repeated structural element and can be made of nearly endless combinations of atoms, compared to the limited nucleotides or amino acids that make up DNA, RNA, and protein. Certain metabolite classes have predictable structures, such as the repeating two-carbon chains present in lipids, which has facilitated the development of LipidBlast for the accurate prediction and identification of lipidomics data (11). Because the number of potential metabolites is unknown, the calculated statistical probability of a correct match may be inaccurate when identifying molecules by comparing to databases (12). It is also difficult to measure all metabolites in biological samples owing to the broad concentration differences between trace and abundant compounds. In serum, for example, diacylglycerols have been detected at picomolar levels while D-glucose and cholesterol are in the micromolar range (13). To capture the chemical diversity of the entire metabolome, multiple methods must be used because no single technique is ideal for the measurement of all small molecules.

The most common metabolomics tools are nuclear magnetic resonance (NMR), gas chromatography–mass spectrometry (GC-MS), and liquid chromatography–MS (LC-MS). Although NMR is regarded as more quantitative and reproducible (14–16), MS-based methods have the advantage of higher sensitivity and can be combined with various chromatographic methods to measure a large diversity of compounds (14). Many types of mass spectrometers have been developed, each with their own advantages. Quadrupole (Q) mass analyzers [including triple-quadrupole instruments for LC-tandem MS (MS/MS)] are highly sensitive and selective for quantification and identification of targeted metabolites, but because of their low mass resolution they are not ideal for unknown characterization. Combining a quadrupole mass selector to a time-of-flight (ToF) mass analyzer in a QToF instrument maintains the selectivity of the quadrupole while also improving the maximum mass resolution to approximately 40,000–60,000. To obtain higher mass resolution for the identification of compounds, Orbitrap mass spectrometers reach a typical resolution of 250,000 up to a maximum of 1,000,000 for ions with m/z less than 300 (17). Fourier transform–ion cyclotron resonance (FT-ICR) mass analyzers still have the highest resolving power, reaching >2,000,000 resolution (18, 19), though these instruments are less often used for metabolomics due to their large size and challenging operation. Coupling chromatographic separations with MS improves detection sensitivity by preventing matrix effects and provides a

second characteristic for compound identification. GC is typically used to resolve and measure volatile nonpolar compounds, though sample derivatization can expand its detection repertoire to more polar compounds. The two major forms of LC used in metabolomics are hydrophilic interaction liquid chromatography (HILIC) (20) and reversed-phase (RP), which most effectively resolve polar and nonpolar compounds, respectively. Some studies have combined GC, HILIC, and RP methods because of their complementary analytes, leading to a comprehensive view of the metabolome (21).

For the simultaneous identification of thousands of compounds in biological samples, GC-MS and LC-MS are the most widely used in metabolomics and are the focus of this review. Many advances in three major areas (instrumentation, experimental design, and computational analysis of data) related to MS-based metabolomics have been made over recent years. This review evaluates the ongoing analytical challenges in metabolomics studies and the strategies related to these three areas that address or resolve these challenges. In particular, developments in the identification of metabolite unknowns and distinguishing metabolites from other small molecules, separation of metabolites by chromatography, and the efficient and accurate sharing of MS data are reviewed herein.

2. METABOLITE IDENTIFICATION

One of the most significant hurdles in metabolomics is the identification of unknown mass spectral features. Owing to the increasing sensitivity and resolution of mass detectors, thousands of features are now analyzed and detected in a single experiment. The vast majority (over 98%) of features in untargeted MS metabolomics experiments cannot be identified by the standard method of querying retention time and accurate mass against an in-house library or searching accurate mass or MS/MS spectral matching to databases (22). Predicted molecular formulas based on the accurate *m/z* and isotopic pattern of MS features can be searched against chemical databases such as PubChem, ChemSpider, the Human Metabolome Database (HMDB), BioCyc, and KEGG (**Figure 1**). Such databases automatically generate accurate masses of all deposited molecules based on atomic composition and therefore contain the most extensive collection of molecular information to query from. However, the limited chemical space of organic metabolites means there are many potential isotopes, which makes identification using accurate masses difficult. Chromatography retention times prior to MS can often distinguish isomers, but deconvolution of coeluting analytes is not always possible when attempting to optimize for thousands of compounds, as in metabolomics. In addition, databases only contain a fraction of existing biomolecules, either due to the compound not yet being identified or because of slow deposition of molecules into databases. The current gold standard for identification involves the use of isotope-labeled internal standards, but these are not always available or economically feasible to purchase or produce. Thus, there has been a strong interest in improving compound identification through methodological, instrumentational, and computational means. Substantial and sustained funding to build, maintain, and make these databases publicly accessible is needed.

2.1. Tandem Mass Spectrometry

The development of MS/MS and its successor, multistage mass spectrometry (MS^n), has improved the ability to distinguish isomeric and isobaric compounds and acquire structural information from accurate masses. MS^n involves the repeated fragmentation and *m/z* determination of ions captured from the previous cycle of MS. Advancements in mass detection sensitivity and ion collection with quadrupole and ion trap technologies have facilitated the extension of fragmentation and mass detection beyond MS/MS. With additional MS cycles, the collision energy can be varied to

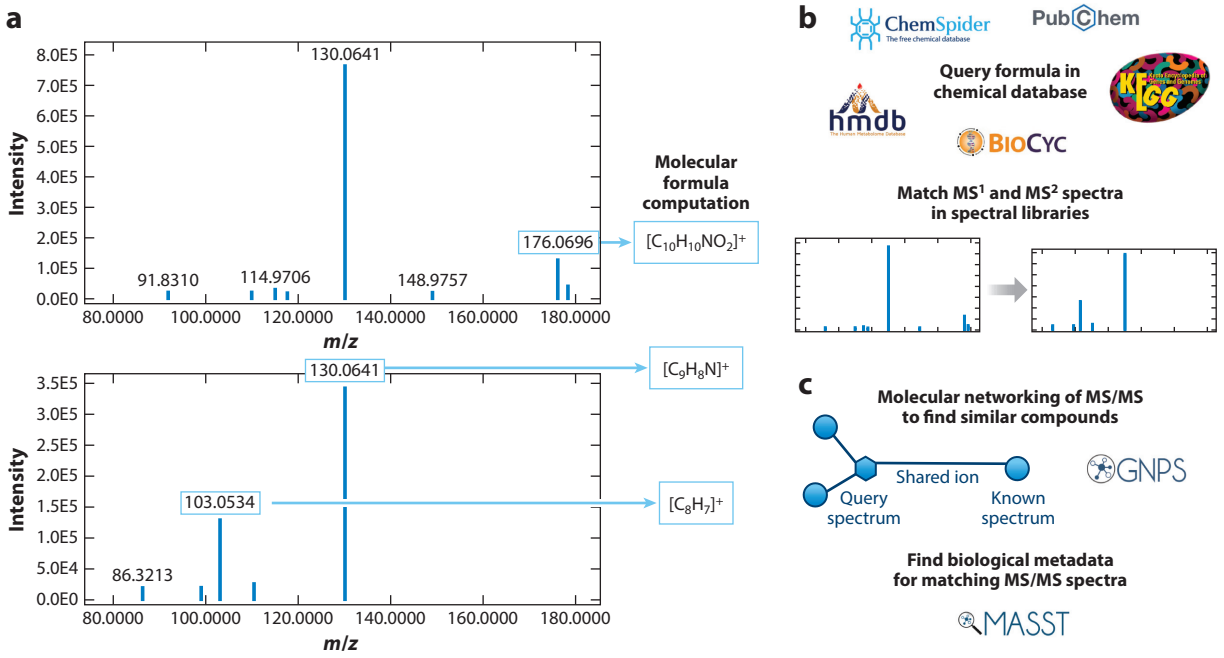

Figure 1

Compound identification strategies using tandem mass spectrometry (MS/MS) spectra. (*a*) Example MS/MS spectra for indole-3-acetic acid (PubChem ID: 802) at different collision energies. Using accurate *m/z* and isotopic abundances, potential molecular formulae of each ion are generated. (*b*) Computed formulae and accurate masses can be searched in chemical databases to find potential candidates. Mass spectra can also be queried against spectra of identified compounds in MS/MS databases, which can match entire spectra or common fragment ions. (*c*) If the compound is a "known unknown," generating a molecular network to compare common fragmentation ions in known compounds can help to classify the molecule. Through searching user-deposited libraries of spectra and their metadata, information can be gained about the biological context of matching unknown spectra.

produce unique subsets of fragmentation ions, allowing for the collection of a richer data set on the substructures and neutral losses originating from a single precursor ion (**Figure 2**).

Many approaches and software have been developed to identify compounds from complex MS^n data. The simplest and most high-throughput technique is to query acquired MS/MS or MS^n spectra against existing MS/MS spectral databases. The most commonly used tandem MS databases for metabolomics are MassBank (23), National Institute of Standards and Technology (NIST), METLIN (24), Global Natural Product Social Molecular Networking (GNPS) (25), mzCloud, HMDB (26), Spektraris (27), and ReSpect (28) (**Figure 1**). Comprehensive reviews of available LC-MS/MS databases already exist (e.g., 12). A notable limitation to metabolite identification through MS/MS databases is the lack of spectra deposited, particularly for increasing cycles of MS^n. Efforts by mzCloud are rapidly expanding the number of spectra beyond MS^2, including the development of detailed spectral trees. Furthermore, of the above MS/MS databases, only mzCloud, NIST, and MassBank report MS^n spectra. Because NIST is not an open-access database, it lacks the benefits of community contribution and data curation present in both mzCloud and MassBank. The unavailability of comprehensive MS^n fragmentation databases therefore makes compound identification more challenging using spectral matching. In addition, retention time, mass spectra, and fragmentation spectra are also dependent on the instrumentation and technique used by the depositor. The type of mass spectrometer, collision

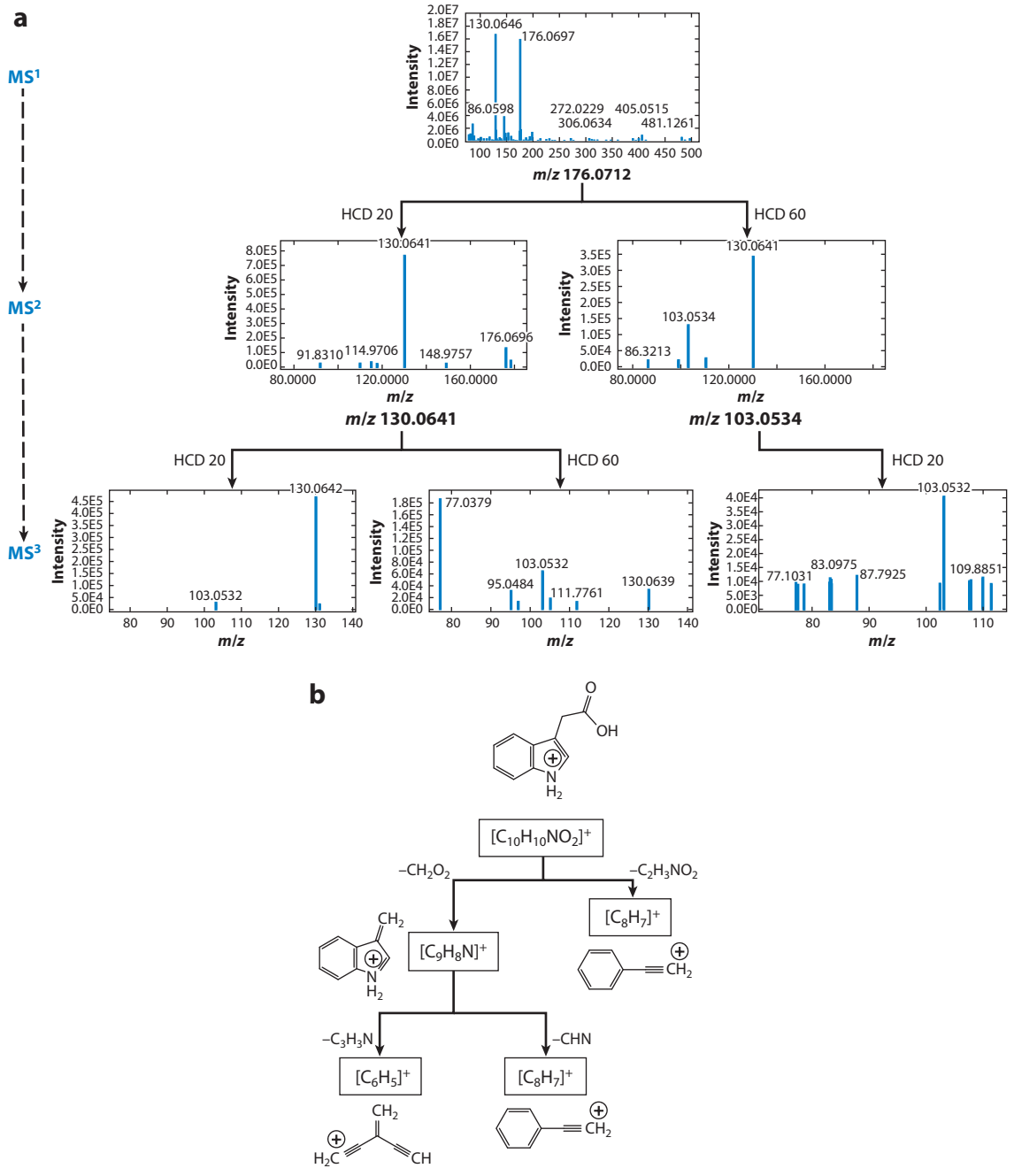

Figure 2

Assembly of the fragmentation tree of indole-3-acetic acid (PubChem ID: 802) from MS^n fragmentation data. (*a*) Specific ions are selected for subsequent rounds of fragmentation and mass spectrometry at varied collision energies (HCD), producing mass spectral trees. (*b*) Fragment ions are translated to molecular formulae by their accurate mass and isotopic abundances, where they are assembled into fragmentation trees, adjoined by the neutral losses between ions. Abbreviations: HCD, higher-energy collisional dissociation; MS^n, multistage mass spectrometry.

energy, collision gas, and pressure influence fragmentation and therefore impede the reliability of comparing MS^n spectra to databases. Chromatographic retention is highly dependent on the method and solvents used, and HILIC methods are especially prone to instability, so database retention times are unreliable for comparison.

To combat these problems, many investigators develop in-house libraries to compile accurate MS and RT information of all known compounds of interest specific to their methods and instrumentation. Tada et al. (29) outlined a method to develop an in-house LC-MS/MS library for all ion fragmentation data-independent acquisition. This method involves running a collection of standard compounds, including standards that have known retention times, to correct for elution variations on the HILIC column (30). Once precursor ions are detected in MS-DIAL (open-access software), their fragmentation spectra are deconvoluted from coeluting analytes using the MS^2Dec and correlation-based deconvolution (CorrDec) algorithms (31). CorrDec identifies fragmentation spectra for each precursor ion by correlating the abundance of MS^1-MS^2 ion pairs across all samples (32). MS^2 ions are then annotated in MS-FINDER using their accurate mass and isotopic pattern, and the MS^2 library is curated and made available to other researchers using the library management tool MS-LIMA. Creating lab-specific libraries is especially advantageous for LC-MS^n because each fragmentation spectrum is dependent on the precursor ions generated in the parent spectrum and many method-specific parameters significantly impact the fragmentation pattern (e.g., collision energy, polarity), leading to compounding differences over subsequent rounds of MS.

2.2. De Novo Compound Identification

Many of the metabolites detected by mass spectrometry are previously unknown compounds and therefore cannot be identified by library searches (33). The utilization of high-resolution mass spectrometers (i.e., ToF, Orbitrap, or FT-ICR analyzers) reduces the number of potential elemental composition matches but does not provide confirmatory identification. This is particularly true for nonmodel organisms and bacterial communities, of which metabolic pathways have not been fully characterized. These "known unknowns" (34) require the use of de novo deconvolution from their MS^n spectra and can open the door for biomarker discovery or illuminate novel metabolites. In the past, de novo structural identification from MS/MS spectra required meticulously honed skill and understanding of the fragmentation patterns of molecules, including common neutral losses and adducts. Along with the complex organization of MS/MS and MS^n data itself, the difficulty of data analysis has initiated the development of numerous computational tools to automate and simplify the process.

The first step of MS/MS analysis typically involves using the accurate mass and isotopic patterns to assign potential molecular formulae to each mass spectral peak, including all fragmentation spectra (**Figure 2**). The spectral peaks and associated molecular formulae are then assembled into fragmentation trees, which link subsequent fragmentation formulae to precursors (nodes) by the neutral losses that differentiate them (edges) (35). Fragmentation trees differ from mass spectral trees that connect mass spectra to the precursor ion(s) selected by the mass spectrometer, which can be used separately for MS/MS spectral matching against libraries (35). The neutral losses calculated from fragmentation trees are then used with the known frequencies of certain neutral losses in the literature and the error in mass accuracy from proposed molecular formulae to rank the most likely formulaic candidates (36).

To designate fragmentation trees with structures of unknown compounds, in silico fragmentation of candidate molecules is performed and matched to experimental spectra (**Figure 3**). Several unique approaches to generating de novo mass spectra from all known chemicals with a

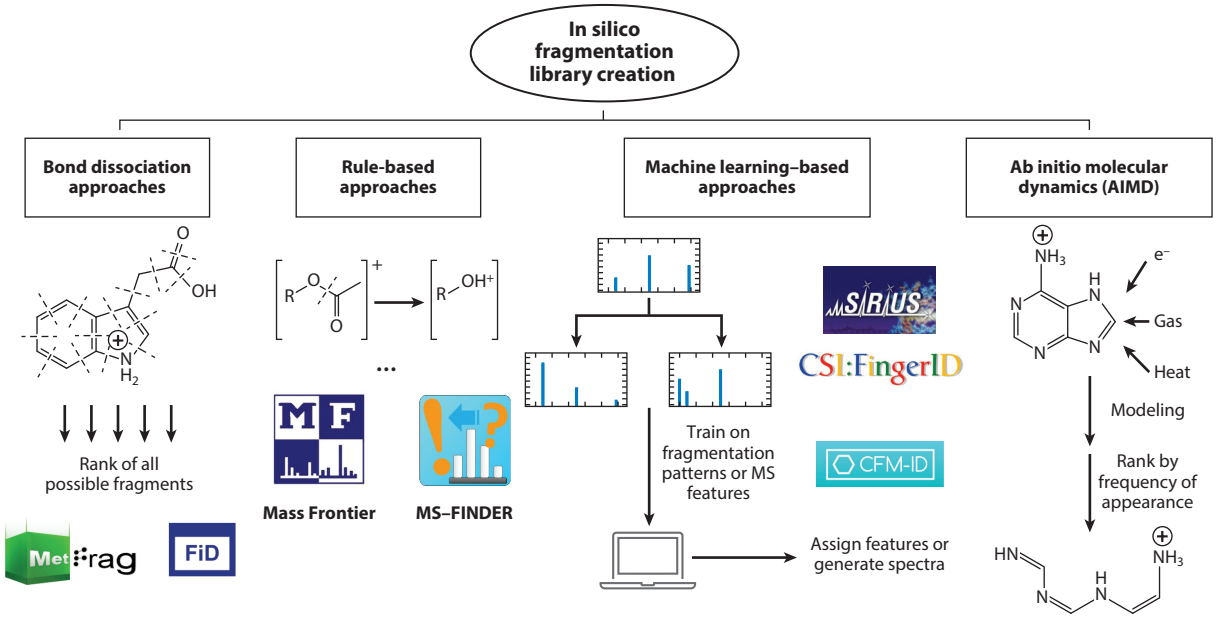

Figure 3

Strategies and software for the in silico generation of fragmentation MS. Bond dissociation programs (e.g., MetFrag and FiD) sequentially break all bonds in the molecule of interest and rank the likelihood of each fragment occurring in the spectrum. Software programs like Mass Frontier and MS-FINDER use their in-house library of chemical reaction rules to predict fragmentation patterns. Machine learning–based approaches train algorithms with existing MS data to generate fragmentation spectra or identify chemical structures from experimental mass spectra. Ab initio molecular dynamics model various types of molecular dissociations brought on by heat or by electron or inert gas collisions and rank the likelihood of generating fragments by how many simulations they appear in. Abbreviations: FiD, Fragment iDentificator; MS, mass spectrometry.

specific molecular formula have been developed into software. The bond dissociation approach uses molecular modeling to sequentially break each bond in the molecule to generate all possible fragmentation ions, then the most likely ions are selected by a weighted scoring metric. MetFrag matches the given accurate mass with molecules in KEGG and PubChem databases to create a candidate list (37). The predicted fragmentation ions are scored in MetFrag by the relative energetic cost of breaking each bond to generate them, then they are assembled into spectra with each ion's relative peak abundance correlating to their relative score (38). A similar approach is also utilized by Fragment iDentificator (FiD) (39). While these tools can predict spectra truly de novo, they are computationally intensive and therefore time consuming.

There has been a great deal of interest in using molecular dynamics (MD) modeling and quantum chemistry to predict molecular fragmentation in silico (40) because of their complete independence from prior knowledge of potential fragmentation patterns (**Figure 3**). These simulations rely on statistical methods (namely Rice–Ramsperger–Kassel–Marcus theory) or MD methods (namely Born–Oppenheimer MD theory) to predict the trajectory of molecular fragmentation upon heating or collision. In the Quantum Chemistry Electron Ionization MS (QCEIMS) program, statistical and MD methods are combined to improve simulations of electron ionization MS (41), and the method has been successfully applied to predict the MS of organic drugs (42) and nucleotides (43, 44). First principles have also been applied to model collision-induced dissociation (CID) of small molecules, including urea (45), galactose-6-sulfate (46), testosterone (47), and sulfated L-cysteine (48). Although modeling often identifies the correct fragmentation ions, it is less

accurate at predicting peak abundances because of the relatively short (picoseconds) time of simulation compared to actual dissociation (milliseconds) (49) and because computational limitations only allow the simulation of single molecule collisions at one time (50). The greatest challenge for modeling-based techniques is the amount of computational time required to predict the bond dissociation of all existing and potential metabolites, particularly as the size of query molecules increases. Thus far, this has not been feasible for generating MS/MS spectra for more than single small molecules, let alone whole MS/MS libraries.

Rule-based tools use a set of curated guidelines to predict the fragmentation of given structures (**Figure 3**). Mass Frontier (Thermo Scientific) uses its large database of individually curated fragmentation rules, such as relative bond strengths, from the literature to predict fragmentation ions. To increase the throughput of the Mass Frontier in silico fragmentation tool, researchers have developed an open-access software package called HAMMER (high-throughput automation of Mass Frontier) to generate predicted MS/MS libraries (51). MS-FINDER software automatically predicts candidate molecular formulae from accurate mass and isotopic patterns and generates all possible fragmentation ions for these candidates (52). Using a weighted ranking from their nine hydrogen rearrangement rules, error in accurate mass, bond dissociation energies, and fragment linkages, the most likely candidate is predicted (53). The advantage of rule-based approaches in comparison to learning-based approaches is that independence from the literature reduces biases in the training set, particularly because existing databases remain incomplete. However, the complexity of metabolite space limits the ability to capture the fragmentation pattern through known rules, as has been done with more regularly structured compounds such as lipids (11).

Others have aimed to improve spectral predictions by training machine learning algorithms with molecular fragmentation data, rather than by producing in silico spectra (54) (**Figure 3**). Competitive Fragmentation Modeling-Identification (CFM-ID) contains a probabilistic generalized modeling algorithm trained on molecular fragmentation patterns in MS/MS data to predict structure from experimental fragmentation spectra (55). CFM-ID also ranks candidate structures from fragmentation spectra using the data from this trained algorithm, such as which neutral losses correspond to substructures in candidates. CSI (Compound Structure Identification):FingerID trains a machine learning algorithm with existing fragmentation trees in MS/MS databases for their pattern of molecular features, the fingerprint (56). It then translates experimental MS/MS data into a fragmentation tree, acquires the molecular fingerprint of the compound, and predicts the structure based on the existence of molecular features in a ranked list (56, 57). The major difference between these machine learning approaches is that CFM-ID trains how molecules preferentially fragment, whereas CSI:FingerID trains how the presence of substructures translates to fragmentation tree features.

While there have been great developments in spectral prediction, many of these techniques still rely on the deposition of the compound in chemical libraries (e.g., KEGG or PubChem) so they are unable to generate complete MS^n spectra. Thus, several tools have been created to help with manual identification or characterization of unknowns. Substructure analysis allows the user to pick out common features in their fragmentation spectra to help assemble or categorize the metabolite of interest. MS2-latent Dirichlet allocation (MS2LDA) and its successor MS2LDA+ identify substructures that they call Mass2Motifs (fragments and neutral losses), based on fragmentation trees, and annotate these substructures in the spectra (58, 59). The NIST hybrid search algorithm allows spectra to be evaluated for the existence of specific substructures based on neutral losses and categorized with these molecules to aid in identification (60). Another approach is to generate molecular networks from MS/MS spectra (**Figure 1**). With this technique, spectral features from both known and unknown compounds (i.e., accurate masses, fragmentation ions, and neutral losses) can be compared to known compounds. Structurally related compounds tend

to form clusters in the network due to them having similar components, so when knowns are included, unknowns can be assigned to chemical groups. This was used by Watrous and others to identify a novel surfactin by its close association in the network with other bacterial surfactins produced by *Bacillus subtilis* (61). Similarly, multivariate analysis of MS/MS data between experimental groups can reveal closely related compounds by their similar reaction to various treatments. For example, orthogonal projections to latent structures discriminant analysis (OPLS-DA) of the urinary metabolome between control and PCN-treated mice identified a cluster of similarly fluctuating metabolites that were determined to be carboxyethyl hydroxychroman (CEHC) metabolites, thus allowing for the identification of the previously unknown γ-CEHC glucuronide (62). Categorizing unknown compounds into biological context with other metabolites can be illuminating for identification. If this cannot be accomplished for an unknown within the existing data set, the mass search tool (MASST) in GNPS is also useful to query MS/MS spectra against the large database of user-submitted MS/MS spectra and associated metadata (63). Knowledge that a certain metabolite is found in a specific tissue, treatment, or organism can provide the necessary context to narrow down potential candidate molecules (**Figure 1**).

Major developments in MS-based metabolomics have focused on improving methods of compound identification from the thousands of features acquired from a single LC-MS or GC-MS run. The widespread use of MS^n and advancement in mass spectrometer technologies have expanded the accuracy and breadth of data collected from compounds present in each sample. However, the most significant recent contributions to compound identification have been the computational methods and tools that parse out meaningful information from mass spectra and programs using artificial intelligence that can predict the structure of unknown compounds. With the ever-growing abundance of information that can be generated from metabolomics experiments, computational tools will continue to be necessary to manage and streamline the compound identification process.

3. DISTINGUISHING TRUE BIOLOGICAL METABOLITES

In complex biological matrices such as plasma, urine, and stool, there are an immense number of compounds from dietary, microbial, host, and other sources. In any given untargeted high-resolution LC-MS experiment, tens to hundreds of thousands of features may be detected, of which less than 10% are true, nonredundant metabolites (64). This continues to be a problem in the metabolomics field, as the critical univariate and multivariate statistics required to identify significant features from the data may be confounded by aberrant features. Use of different ionization modes to expand the chemical diversity captured in an experiment also results in redundant compounds. Given the difficulty and time requirement of compound identification, there is a great need to remove undesired features in metabolomics data.

To some extent, feature reduction has become a standard approach (**Figure 4**). Combining compound adducts, which involves the identification and joining of MS peaks formed from different adduct ions (e.g., $[M+H]^+$, $[M+Na]^+$, $[M+NH_4]^+$), has been incorporated into software such as MS-FLO (MS–feature list optimizer) (65). Different adducts form depending on the composition of ions and other compounds in the sample matrix. MS-FLO also removes isotopes from MS data based on the accurate mass shift and expected peak height ratios between isotopes, usually ^{13}C, ^{2}H, ^{15}N, and ^{18}O. There has been recent interest in combining analytes from different ionization modes (positive and negative) to streamline the number of features to be identified. MSCombine was developed to address this problem, with success in human serum and urine data sets (66). The simple adducts that are used to combine spectral peaks in most feature reduction software represent a small fraction of the adducts formed by a single metabolite, which can be upward of

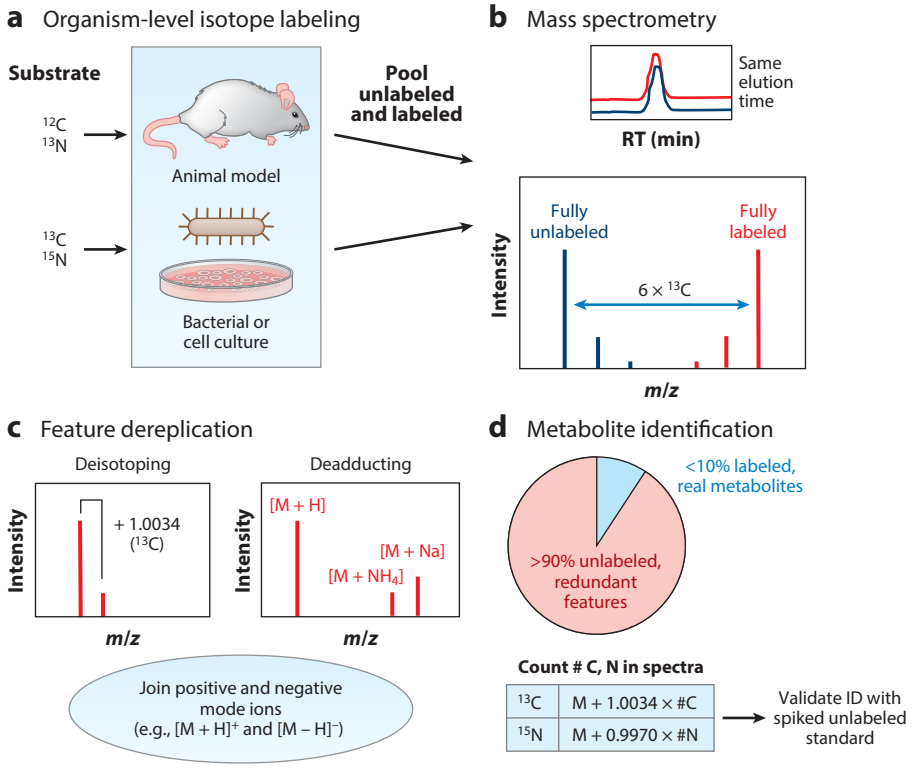

Figure 4

Isotope labeling and feature dereplication to distinguish true metabolites. (*a*) Model organisms are labeled with ^{13}C and/or ^{15}N substrate, and a portion of each metabolite extract is pooled and analyzed with high-resolution mass spectrometry. (*b*) Labeled features are identified by a greater *m/z* at the same elution time corresponding to the exact mass difference of additional heavy isotopes (i.e., ^{13}C and ^{15}C). In the case of isotopic ratio outlier analysis (IROA), a U-shaped isotopologue pattern occurs between the completely unlabeled and labeled species. (*c*) Redundant features such as naturally occurring isotopes, adducts, and duplicate features from positive and negative mode analyses are also dereplicated from the data. (*d*) Filtering of isotope-labeled features and dereplication reduce features to less than 10%. Whole-organism isotope labeling can also assist in identifying features by counting the carbon and nitrogen atoms between fully labeled and unlabeled spectral peaks, which can be validated by spiking unlabeled standard compounds with fully labeled organism extracts. Abbreviation: RT, retention time.

100 features for a single compound run by LC-MS (67). To recognize and merge more complex adducts, including those made of multiple analytes, the software mz.unity generates all possible adducts based on the accurate mass of the neutral loss between any two spectral peaks within a specified mass error (67). These techniques ensure that each identified compound only appears once, thus preventing the overweighting of compound duplicates in downstream statistical analyses.

Other spectral artifacts or contaminants that are not metabolites of the organism of interest are much more challenging to recognize. Stable isotope–fed organisms can be used distinguish the metabolites they produce from environmental chemicals (**Figure 4**). When model organisms are fed isotope-labeled substrate, true metabolites incorporate the heavy isotopes and are observable in MS by a specific increase in the *m/z* while retaining the same elution properties (68). Isotopic ratio outlier analysis (IROA) uses the isotopic labeling pattern of cells grown in ^{13}C-glucose to

identify metabolites from mass spectra (69–71). By pooling extracts from individual treatments grown in either 5% or 95% ^{13}C, an identifiable U-shaped isotopologue pattern is produced for each metabolite, corresponding to the incorporation of ^{13}C into each carbon in the molecule. Heavy isotope labeling allows metabolites to be distinguished from background ions, and the fully labeled isotopologue can help to ascertain the number of labeled species in the analyte. Mahieu & Patti (64) applied both degenerate feature reduction and organism-level ^{13}C-glucose labeling to LC-MS metabolomics of *Escherichia coli* and reduced the total number of features from over 25,000 to fewer than 1,000. Wang et al. (72) similarly developed the Peak Annotation and Verification Engine (PAVE) to remove redundant MS adducts and identify metabolites from isotope-labeled organisms (**Figure 4**). However, they made several improvements, including using both ^{13}C- and ^{15}N-labeled substrates to identify metabolite ions and incorporating a weak collision step to help differentiate the parent ion within spectra (72). Using heavy isotope incorporation to narrow feature lists to only include metabolites reduces the time investment dedicated to identifying compounds and ensures that the features remaining are not environmental contaminants. This technique has widespread applicability to bacteria, plants, and other model organisms but is limited in the types of experiments it may be used in. Treatments or genetic manipulations that alter the primary carbon source used by the organism will interfere with labeling in that condition. Owing to the requirement for administering labeled substrate, this technique is also not applicable to exposome studies that measure exposure to diet or environmental chemicals. Still, true metabolite feature libraries may be catalogued to screen compounds from future experiments involving the organism.

4. CHROMATOGRAPHIC SEPARATION OF METABOLITES

When performing MS-based metabolomics on heterogeneous biological samples, chromatographic separations such as GC or LC are generally used prior to MS. Chromatography is especially valuable owing to its ability to distinguish isomers by their physical properties, which remains a major drawback of direct infusion MS. The most commonly used types of LC columns in metabolomics are RP and HILIC, in which compounds interact with a nonpolar hydrocarbon solid phase or an aqueous layer supported by a polar solid phase, respectively. Metabolite quantification is therefore related to peak area or height in chromatographs and either used as is for relative abundances or compared to a calibration curve to obtain absolute values. Coupling MS to chromatography provides an additional level of certainty when identifying compounds because elution order is a relatively stable property of molecules. According to the Metabolomics Standards Initiative Chemical Analysis Working Group (CAWG), retention time is an effective orthogonal property for metabolite identification, level 1 (73).

4.1. Retention Indices and Retention Time Prediction

Retention times are highly susceptible to variation based on a variety of factors, including the column and instrument used, solvent preparation, flow rate, temperature, and solvent gradient parameters. This has limited the development and use of retention time libraries as a method of metabolite identification, as any deviation from the reference method may render retention time values inaccurate. Ideally, pure standards run by the user with the same method and instrument are compared to the retention time observed in the sample. However, standards are not available for most known unknowns and synthesis is often expensive and time consuming. HILIC methods are especially susceptible to variations in retention time from small changes to buffer composition, even when using identical methods and instrumentation (30). As a way to normalize retention time differences between batches, instruments, methods, labs, and even from different sample types

(urine, plasma, stool) (74), retention indices for GC separations were introduced by Kováts in 1958 (75). The retention index is a set of retention time values normalized to standard compounds that elute at semiregular intervals through the chromatograph. Ideal retention markers span the entire chromatogram to account for all measured compounds and are chemically alike to the compounds being measured to ensure they are affected similarly by chromatographic conditions. Here, we establish the current state of GC retention indices and how retention prediction has been recently applied to LC.

Linear retention indices were originally designed for GC and continue to be effective and widely used for GC retention time correction due to the relatively consistent correlation between a compound's retention index on a GC column and its boiling point. Additional GC indices have been developed using different marker compounds and applied to isothermic and temperature-programming GC conditions. The most common linear indices for GC analysis are the Kováts index, using *n*-alkanes differing by single linear hydrocarbons (75, 76), and the Lee index, using polycyclic aromatic hydrocarbons (77). Several other indices have been developed for specific classes of compounds, such as the fatty acid methyl esters (FAMEs) created for GC lipid analyses by the Fiehn lab (78). FAMEs are better than *n*-alkanes at retaining their molecular ion and generating distinguishing fragmentation ions during electron ionization, making them an improved retention index compared with the Kováts index for GC-MS methods (74). Software tools in the NIST database have been developed for the automated calculation and analysis of Kováts, linear, and Lee retention index values (79), making retention index values for in-house compounds easy to acquire. Retention index values can also be compared to literature values through NIST or other metabolite databases that report retention index values such as the Golm Metabolome Database (80) or MassBank (23). Because not all compounds are represented in these databases, algorithms to predict the retention of unknowns in relation to indices are valuable to aid in identification. Currently this can be done to predict Kováts index values for unknowns based on the presence of various chemical groups (e.g., hydrocarbon chain length, carboxyls, carbonyls) (81).

Although linear retention indices have been effective for GC retention time correction, their application to LC methods has proven to be more unreliable for several reasons. Compound elution times in LC columns are challenging to predict because they do not always follow linear relationships with the index, sometimes even switching their elution order with other compounds. Rather than using a linear retention index, retention projection identifies isocratic retention factors for each compound based on experimental elution values. These retention factors indicate the specific solvent properties in the gradient where the compound elutes, which are then used to predict its retention in a new gradient (82). One problem with this method is that instrumental error or inaccurate solvent preparation can cause the actual gradient to differ from what the instrument reports (83). However, labs can combat this by running standards with known elution properties, calculating the actual gradient and correcting retention projection values according to the error observed. This has been shown to effectively correct retention time deviation caused by the use of different instruments, gradients, and flow rates (84) and even between labs (82), thereby reducing error compared to linear retention indices by more than half.

Predictive algorithms have been developed for the estimation of the retention time of unknowns. This approach was initiated by using basic structural determinants of polarity and charge and the elution properties of known compounds to predict retention of unknowns with similar chemical structures, though these values have error (85). Recent advances in quantitative structure-retention relationship (QSRR) models have been able to improve these predictions in both RP and HILIC LC applications (86). QSRR uses machine learning to increase the accuracy of identifying relationships between structural features and polarity by training on real data sets, which can

then be used to predict how they elute in a column. One of the major advantages to this method is that the use and validation with real data facilitate accurate estimations of error in retention time prediction, which can be used to better reject potential compound identities (87). In addition to physicochemical properties, other molecular descriptors (i.e., fingerprints) have been incorporated in a kernel-based partial least squares model to improve retention time predictions (88). Machine learning techniques require training from data sets containing accurate and extensive retention time data from compounds that are chemically similar to unknowns. This has been a major limitation because certain classes of molecules are difficult to acquire or synthesize and accurate predictions are best performed by comparison to in-house libraries using the same column, instrument, and gradient.

Despite the limited availability of retention time data sets and the sensitivity of chromatography to fluctuation, retention predictions can be valuable when coupled with MS. The accurate mass and isotopic pattern of unknown compounds determined by MS can be used to predict the molecular formula and therefore all possible chemical structures. Although retention time predictions can only be used to reject candidate structures, this could significantly narrow down potential candidates to a manageable number to be validated by running standards. With further improvements in retention library size, availability of diverse chemicals for purchase, and accuracy in machine learning algorithms, the retention time error window will continue to be narrowed.

4.2. Two-Dimensional Chromatography

Metabolomics analyses on complex biological matrices often demand the identification of chemically diverse compounds, ideally using as few separate methods as possible. A major issue with typical chromatographic separation is that each method is optimized for separating only compounds that are retained on the column. This is especially true in LC; for example, polar compounds run by RP will all elute nearly simultaneously within the first few seconds because of their strong preference for the polar solvent over the column. Poor separation makes quantification less accurate and lowers the sensitivity due to matrix effects. To improve the separation of additional compounds from one sample, two-dimensional (2D) chromatography can be used. For LC applications, the most common combination is RP and HILIC because of their high orthogonality, meaning they are able to resolve complementary compounds (89, 90). However, when attempting to resolve highly similar compounds, RP-RP has also been used. A study by Willmann et al. (91) used RP-RP-MS to quantify RNA metabolites for cancer biomarker identification.

There are several approaches to LC-LC that manage the challenge of the second dimension taking time to run while molecules are continuously eluted from the first column. Heart-cutting chromatography is a form of 2D chromatography that targets specific peaks by cutting them from the spectrum and directing them to the second dimension. Extracting individual peaks facilitates high-resolution separation and the data are relatively easy to analyze, though it only acquires LC-LC for few selected compounds. Comprehensive LC-LC (or LC×LC), which is of major interest for metabolomics, collects secondary LC data for the entire first chromatogram. However, the incompatibility of solvents between the first and second rounds of chromatography often results in peak splitting and lowers the sensitivity and resolution of the second chromatograph (92). In a recent study, 100 lipids were identified in bovine urine using comprehensive RP-HILIC-MS (93). The team addressed the solvent incompatibility issue by diluting the first-dimension eluate with the proper infusion solvent for the second dimension, but this substantially lowered the sensitivity of detection. Others have used vacuum collection or stop-flow using a trap column after the first separation to extract analytes from solvent before redissolving in the appropriate injection solvent

for the second column (94). Stop-flow LC×LC was developed by Wang et al. (95), who used RP-HILIC-MS to identify an additional 88 lipids in human plasma compared to RP alone, while maintaining good linearity for quantification.

Although the technology for LC×LC acquisition and analysis continues to improve, there are several barriers preventing its widespread applicability to metabolomics research. There is currently no effective data analysis software to organize and analyze complex 2D chromatography data, especially because the technique is often coupled with MS. Due to the numerous eluent collectors, columns, additional tubing, and flow switches, the instrumentation itself is expensive and challenging to upkeep. In addition, each run requires more time for the separation of a fraction more analytes, and method development is difficult and time consuming. As an alternative to 2D chromatography, developments in computational methods for LC and GC peak deconvolution have become an increasingly viable alternative for distinguishing peaks. Historically, MS^1 spectra from overlapping chromatographic peaks have been used to ascertain the relative abundance of coeluting compounds. For isotopes with similar MS^1 spectra, recent methods have been developed to deconvolute based on MS^2, such as in MS-DIAL (31). Given the significant time and analysis challenges and available alternatives, large advances must be made for the widespread application of 2D chromatography in metabolomics.

5. MASS SPECTRAL DATABASES AND COMMUNITY SHARING

Proper identification of mass spectral features and classification into metabolic pathways require the use of high-quality curated MS databases, such as METLIN, mzCloud, NIST, MassBank, HMDB, or ReSpect. Classic compound identification relies on the comparison of characteristic ion *m/z* abundances to those of known compounds and the scoring of potential candidates based on their similarity. Machine learning algorithms for the in silico prediction of mass spectra and assignment of substructures are also refined by larger training sets from various MS databases. However, of the known compounds that are listed in chemical databases like PubChem (>100 million), only a small fraction has MS^1 spectra in these MS databases, with even fewer MS^n spectra present. Nonetheless, with the increasing deposition of spectra into databases, Mass Frontier and other programs have been able to develop tools to synthesize MS^n data and generate spectral trees for improved analysis. For a thorough review of available MS databases and the types of retention time, mass spectral, and tandem mass spectral data they house, see recent reviews (96, 97). The purpose of this section is to highlight recent advances in mass spectral libraries and what remains to be done to improve the field of metabolomics research.

One of the major problems beyond a simple lack of data is that the compounds in these spectral databases do not evenly represent all metabolites found in nature. Compounds that are difficult to ionize or that are low in abundance in biological samples often do not have acquired spectra. Furthermore, many small molecules remain completely unknown, particularly those produced by nonmodel organisms, including most bacteria, and therefore do not have associated mass spectra in databases. Although complete in silico generation of mass spectra is possible to putatively identify metabolites, the accuracy of these predictions is dependent on chemically similar metabolites being present in the training database (39, 54, 56). The spectra present in databases are also biased toward those acquired by GC-MS (i.e., more volatile metabolites) because of its earlier development and easier compound identification compared to LC-MS. However, LC-MS is often preferred in metabolomics to capture a larger diversity of biomolecules. Since the instrument, collision energy, and other parameters between labs and experiments can affect the spectra generated, compound identification requires that mass spectra in databases come from similar instruments and protocols. It is of utmost importance that annotated mass spectra from all types of instruments

and newly discovered compounds continue to be deposited into MS and chemical databases to increase the percentage of features that are identified in metabolomics experiments.

Biological samples often contain a mixture of metabolites originating from multiple organisms, including plants, mammals, and bacteria. Human feces, for example, contain human metabolites, indigestible small molecules from the diet, xenobiotics, gut microbiome metabolites, and metabolites from the combined metabolism of host and microbes (98). It can therefore be challenging to identify and distinguish the source of metabolites without extensive knowledge of the types of reactions these organisms perform. There has been a recent surge in databases that use the known metagenomes of humans and bacteria to predict the source of identified metabolites and their potential contribution to human disease, including the Virtual Metabolic Human (VMH) database (99) and AMON (100). However, owing to the lack of known bacterial products and extensive curation and metadata required to accurately classify compounds in these databases, they only contain very few (~5,200 in VMH) metabolites.

Data sharing in metabolomics from a variety of contributors is essential to impart context to how, where, and when certain metabolites are produced by living systems. In 2007, the Metabolomics Standards Initiative recommended the creation of a central repository for metabolomics data sets with strict guidelines for submission (101). One of the major problems with user-submitted data is inconsistent naming, which prevents the appropriate automatic compilation of data from multiple sources that would be impossible to curate manually (102). Many have suggested that users submit all data with InChI and PubChem IDs because of their universal, nonoverlapping, and machine-readable naming systems (96, 103, 104). Kind et al. (104) have also recommended that all newly identified compounds be automatically submitted to community libraries with extensive metadata related to the sample preparation (organism or sample source, extraction steps, time course, etc.) and data collection (instrument, column, gradient, collision energy, etc.). GNPS (25), the Metabolomics Workbench (105), and MetaboLights (106) are the major MS data sharing platforms (12) that house a collection of mass spectra, metadata, protocols, and known metabolite structures. Of these, GNPS has many additional functions that use its community data repository Mass Spectrometry Interactive Virtual Environment (MassIVE) to make the analysis of user-submitted spectra and metadata particularly approachable and useful for researchers (25). MassIVE is a living library of user-submitted MS data that are continuously curated by researchers and GNPS to identify compounds as their spectra become known. Thus, previously submitted data to MassIVE will be updated with the identity of compounds and substructures based on the newest annotations. The MASST search function in GNPS allows researchers to find what other sample types contain similar unidentified spectra based on their associated metadata, which can give clues to the identity of the unknown (107). Although increased data sharing in user-submitted libraries is ultimately beneficial, it is important to extensively and accurately curate data. Errors in submitted chemical identifiers and spectral peak intensities due to coelution of peaks can greatly impact compound identifications with libraries like GNPS that use these data to annotate other data sets (102). Moving forward, the deposition of mass spectra from validated standard compounds run on a variety of instruments will become increasingly important to improve not only the size but also the quality of data sharing in metabolomics.

6. CONCLUSION

Through recent advancements in instrument technology, experimental techniques, and analysis software, many of the challenges with MS-based metabolomics have been mitigated. The task of identifying unknown compounds from mass spectra has been improved by the structural information acquired from additional cycles of fragmentation and MS (MS^n), numerous softwares

that have been developed to generate increasingly accurate in silico fragmentation spectra, and the machine learning tools that use real MS data to predict features or candidate structures from experimental MS. For example, by analyzing plant extracts with LC-MS/MS and combining multivariate statistical analysis in MetaboAnalyst and molecular networking in GNPS, Demarque et al. (108) could prioritize features in the data that appeared to impact bioactivity and identified the active compounds to be acetogenins. Although MS data sharing continues to expand, the lack of high-quality spectral data from a diversity of instruments and collision energies remains one of the largest difficulties for metabolite identification. Distinguishing metabolites or important features from environmental small molecules has been another challenge in metabolomics. Computational feature dereplication and the use of isotope-labeled organisms to differentiate true metabolites from contaminants can reduce the number of important features by more than tenfold. However, additional strategies are needed to filter undesired features from exposome studies and samples taken from human subjects that cannot be isotope labeled. The ongoing issue of inconsistent elution from chromatography has been improved with retention indices, which help to normalize retention time values across instruments and enable the use of retention time databases for compound identification. Future funding should be applied to techniques that remove unimportant MS features, expand MS databases and data curation, and automate metabolite identification. In addition, the centralization of data deposition, curation, and computational spectrum analysis, as well as the standardization of analytical methods across various platforms, will improve the confidence of compound identification from MS-based datasets. With this, biologically relevant metabolites will continue to be elucidated for drug discovery, diagnostics, and other applications.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Newgard CB. 2017. Metabolomics and metabolic diseases: Where do we stand? *Cell Metab.* 25:43–56
2. Almontashiri NAM, Zha L, Young K, Law T, Kellogg MD, et al. 2020. Clinical validation of targeted and untargeted metabolomics testing for genetic disorders: a 3 year comparative study. *Sci. Rep.* 10:9382
3. Smolinska A, Blanchet L, Buydens LMC, Wijmenga SS. 2012. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Anal. Chim. Acta* 750:82–97
4. Wang X, Liu L, Zhang W, Zhang J, Du X, et al. 2017. Serum metabolome biomarkers associate low-level environmental perfluorinated compound exposure with oxidative/nitrosative stress in humans. *Environ. Pollut.* 229:168–76
5. Sharon G, Garg N, Debelius J, Knight R, Dorrestein PC, Mazmanian SK. 2014. Specialized metabolites from the microbiome in health and disease. *Cell Metab.* 20(5):719–30
6. Ortmayr K, Dubuis S, Zampieri M. 2019. Metabolic profiling of cancer cells reveals genome-wide crosstalk between transcriptional regulators and metabolism. *Nat. Commun.* 10(1):1841
7. Zenobi R. 2013. Single-cell metabolomics: analytical and biological perspectives. *Science* 342:1243259
8. Johnson CH, Ivanisevic J, Siuzdak G. 2016. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* 17:451–59
9. Zampieri M, Zimmermann M, Claassen M, Sauer U. 2017. Nontargeted metabolomics reveals the multilevel response to antibiotic perturbations. *Cell Rep.* 19(6):1214–28
10. Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28(1):27–30
11. Kind T, Liu K-H, Yup Lee D, DeFelice B, Meissen JK, Fiehn O. 2013. LipidBlast *in silico* tandem mass spectrometry database for lipid identification. *Nat. Methods* 10(8):755–58

12. Blaženović I, Kind T, Ji J, Fiehn O. 2018. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* 8(2):31
13. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, et al. 2011. The human serum metabolome. *PLOS ONE* 6(2):e16957
14. Emwas AHM. 2015. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods Mol. Biol.* 1277:161–93
15. Markley JL, Brüschweiler R, Edison AS, Eghbalnia HR, Powers R, et al. 2017. The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* 43:34–40
16. Gowda GAN, Gowda YN, Raftery D. 2015. Expanding the limits of human blood metabolite quantitation using NMR spectroscopy. *Anal. Chem.* 87:706–15
17. Denisov E, Damoc E, Lange O, Makarov A. 2012. Orbitrap mass spectrometry with resolving powers above 1,000,000. *Int. J. Mass Spectrom.* 325–327:80–85
18. Hendrickson CL, Quinn JP, Kaiser NK, Smith DF, Blakney GT, et al. 2015. 21 Tesla Fourier transform ion cyclotron resonance mass spectrometer: a national resource for ultrahigh resolution mass analysis. *J. Am. Soc. Mass Spectrom.* 26(9):1626–32
19. Shaw JB, Lin T-Y, Leach FE, Tolmachev AV, Tolić N, et al. 2016. 21 Tesla Fourier transform ion cyclotron resonance mass spectrometer greatly expands mass spectrometry toolbox. *J. Am. Soc. Mass Spectrom.* 27(12):1929–36
20. Buszewski B, Noga S. 2012. Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique. *Anal. Bioanal. Chem.* 402(1):231–47
21. Kind T, Tolstikov V, Fiehn O, Weiss RH. 2007. A comprehensive urinary metabolomic approach for identifying kidney cancer. *Anal. Biochem.* 363(2):185–95
22. da Silva RR, Dorrestein PC, Quinn RA. 2015. Illuminating the dark matter in metabolomics. *PNAS* 112(41):12549–50
23. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, et al. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* 45(7):703–14
24. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, et al. 2018. METLIN: a technology platform for identifying knowns and unknowns. *Anal. Chem.* 90(5):3156–64
25. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, et al. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* 34(8):828–37
26. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, et al. 2007. HMDB: the human metabolome database. *Nucleic Acids Res.* 35:D521–26
27. Cuthbertson DJ, Johnson SR, Piljac-Žegarac J, Kappel J, Schäfer S, et al. 2013. Accurate mass–time tag library for LC/MS-based metabolite profiling of medicinal plants. *Phytochemistry* 91:187–97
28. Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, et al. 2012. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry* 82:38–45
29. Tada I, Tsugawa H, Meister I, Zhang P, Shu R, et al. 2019. Creating a reliable mass spectral-retention time library for all ion fragmentation-based metabolomics. *Metabolites* 9(11):251
30. Zhu Q-F, Zhang T-Y, Qin L-L, Li X-M, Zheng S-J, Feng Y-Q. 2019. Method to calculate the retention index in hydrophilic interaction liquid chromatography using normal fatty acid derivatives as calibrants. *Anal. Chem.* 91(9):6057–63
31. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, et al. 2015. MS-DIAL: data independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* 12(6):523–26
32. Tada I, Chaleckis R, Tsugawa H, Meister I, Zhang P, et al. 2020. Correlation-based deconvolution (CorrDec) to generate high-quality MS2 spectra from data-independent acquisition in multisample studies. *Anal. Chem.* 92(16):11310–17
33. Junot C, Fenaille F, Colsch B, Bécher F. 2014. High resolution mass spectrometry based techniques at the crossroads of metabolic pathways. *Mass Spectrom. Rev.* 33(6):471–500
34. Little JL, Cleven CD, Brown SD. 2011. Identification of "known unknowns" utilizing accurate mass data and chemical abstracts service databases. *J. Am. Soc. Mass Spectrom.* 22(2):348–59

35. Vaniya A, Fiehn O. 2015. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends Anal. Chem.* 69:52–61
36. Böcker S, Rasche F. 2008. Towards *de novo* identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* 24(16):i49–55
37. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. 2016. MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation. *J. Cheminformatics* 8:3
38. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. 2010. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinform.* 11:148
39. Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, et al. 2008. FiD: a software for *ab initio* structural identification of product ions from tandem mass spectrometric data. *Rapid Commun. Mass Spectrom.* 22(19):3043–52
40. Bartlett RJ, Musiał M. 2007. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* 79(1):291–352
41. Bauer CA, Grimme S. 2016. How to compute electron ionization mass spectra from first principles. *J. Phys. Chem. A* 120(21):3755–66
42. Bauer CA, Grimme S. 2014. First principles calculation of electron ionization mass spectra for selected organic drug molecules. *Org. Biomol. Chem.* 12(43):8737–44
43. Bauer CA, Grimme S. 2014. Elucidation of electron ionization induced fragmentations of adenine by semiempirical and density functional molecular dynamics. *J. Phys. Chem. A* 118(49):11479–84
44. Bauer CA, Grimme S. 2015. Automated quantum chemistry based molecular dynamics simulations of electron ionization induced fragmentations of the nucleobases uracil, thymine, cytosine, and guanine. *Eur. J. Mass Spectrom.* 21:125–40
45. Spezia R, Salpin J-Y, Gaigeot M-P, Hase WL, Song K. 2009. Protonated urea collision-induced dissociation. Comparison of experiments and chemical dynamics simulations. *J. Phys. Chem. A* 113(50):13853–62
46. Ortiz D, Salpin J-Y, Song K, Spezia R. 2014. Galactose-6-sulfate collision induced dissociation using QM+MM chemical dynamics simulations and ESI-MS/MS experiments. *Int. J. Mass Spectrom.* 358:25–35
47. Lee G, Park E, Chung H, Jeanvoine Y, Song K, Spezia R. 2016. Gas phase fragmentation mechanisms of protonated testosterone as revealed by chemical dynamics simulations. *Int. J. Mass Spectrom.* 407:40–50
48. Macaluso V, Scuderi D, Crestoni ME, Fornarini S, Corinti D, et al. 2019. L-cysteine modified by S-sulfation: consequence on fragmentation processes elucidated by tandem mass spectrometry and chemical dynamics simulations. *J. Phys. Chem. A* 123(17):3685–96
49. Molina ER, Salpin J-Y, Spezia R, Martínez-Núñez E. 2016. On the gas phase fragmentation of protonated uracil: a statistical perspective. *Phys. Chem. Chem. Phys.* 18(22):14980–90
50. Martin Somer A, Macaluso V, Barnes GL, Yang L, Pratihar S, et al. 2020. Role of chemical dynamics simulations in mass spectrometry studies of collision-induced dissociation and collisions of biological ions with organic surfaces. *J. Am. Soc. Mass Spectrom.* 31(1):2–24
51. Zhou J, Weber RJM, Allwood JW, Mistrik R, Zhu Z, et al. 2014. HAMMER: automated operation of mass frontier to construct in silico mass spectral fragmentation libraries. *Bioinformatics* 30(4):581–83
52. Vaniya A, Samra SN, Palazoglu M, Tsugawa H, Fiehn O. 2017. Using MS-FINDER for identifying 19 natural products in the CASMI 2016 contest. *Phytochem. Lett.* 21:306–12
53. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, et al. 2016. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal. Chem.* 88(16):7946–58
54. Nguyen DH, Nguyen CH, Mamitsuka H. 2019. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief. Bioinform.* 20(6):2028–43
55. Allen F, Greiner R, Wishart D. 2015. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* 11(1):98–110
56. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *PNAS* 112(41):12580–85
57. Ludwig M, Dührkop K, Böcker S. 2018. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics* 34(13):i333–40

58. van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S. 2016. Topic modeling for untargeted substructure exploration in metabolomics. *PNAS* 113(48):13738–43
59. van der Hooft JJJ, Wandy J, Young F, Padmanabhan S, Gerasimidis K, et al. 2017. Unsupervised discovery and comparison of structural families across multiple samples in untargeted metabolomics. *Anal. Chem.* 89(14):7569–77
60. Simón-Manso Y, Marupaka R, Yan X, Liang Y, Telu KH, et al. 2019. Mass spectrometry fingerprints of small-molecule metabolites in biofluids: building a spectral library of recurrent spectra for urine analysis. *Anal. Chem.* 91(18):12021–29
61. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, et al. 2012. Mass spectral molecular networking of living microbial colonies. *PNAS* 109(26):E1743–52
62. Cho J-Y, Kang DW, Ma X, Ahn S-H, Krausz KW, et al. 2009. Metabolomics reveals a novel vitamin E metabolite and attenuated vitamin E metabolism upon PXR activation. *J. Lipid Res.* 50(5):924–37
63. Wang M, Jarmusch AK, Vargas F, Aksenov AA, Gauglitz JM, et al. 2020. Mass spectrometry searches using MASST. *Nat. Biotechnol.* 38:19–26
64. Mahieu NG, Patti GJ. 2017. Systems-level annotation of a metabolomics data set reduces 25000 features to fewer than 1000 unique metabolites. *Anal. Chem.* 89(19):10397–406
65. DeFelice BC, Mehta SS, Samra S, Čajka T, Wancewicz B, et al. 2017. Mass spectral feature list optimizer (MS-FLO): a tool to minimize false positive peak reports in untargeted liquid chromatography-mass spectroscopy (LC-MS) data processing. *Anal. Chem.* 89(6):3250–55
66. Calderón-Santiago M, Fernández-Peralbo MA, Priego-Capote F, Luque de Castro MD. 2016. MSCombine: a tool for merging untargeted metabolomic data from high-resolution mass spectrometry in the positive and negative ionization modes. *Metabolomics* 12(3):43
67. Mahieu NG, Spalding JL, Gelman SJ, Patti GJ. 2016. Defining and detecting complex peak relationships in mass spectral data: the mz.unity algorithm. *Anal. Chem.* 88(18):9037–46
68. Tsugawa H, Nakabayashi R, Mori T, Yamada Y, Takahashi M, et al. 2019. A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nat. Methods* 16(4):295–98
69. de Jong FA, Beecher C. 2012. Addressing the current bottlenecks of metabolomics: isotopic Ratio Outlier Analysis^{TM}, an isotopic-labeling technique for accurate biochemical profiling. *Bioanalysis* 4(18):2303–14
70. Qiu Y, Moir R, Willis I, Beecher C, Tsai Y-H, et al. 2016. Isotopic ratio outlier analysis of the *S. cerevisiae* metabolome using accurate mass gas chromatography/time-of-flight mass spectrometry: a new method for discovery. *Anal. Chem.* 88(5):2747–54
71. Qiu Y, Moir RD, Willis IM, Seethapathy S, Biniakewitz RC, Kurland IJ. 2018. Enhanced isotopic ratio outlier analysis (IROA) peak detection and identification with ultra-high resolution GC-Orbitrap/MS: potential application for investigation of model organism metabolomes. *Metabolites* 8(1):9
72. Wang L, Xing X, Chen L, Yang L, Su X, et al. 2019. Peak annotation and verification engine for untargeted LC-MS metabolomics. *Anal. Chem.* 91(3):1838–46
73. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, et al. 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 3(3):211–21
74. Fiehn O. 2016. Metabolomics by gas chromatography-mass spectrometry: the combination of targeted and untargeted profiling. *Curr. Protoc. Mol. Biol.* 114:30.4.1–32
75. Kováts E. 1958. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helv. Chim. Acta* 41(7):1915–32
76. Kováts E. 1965. Gas chromatographic characterization of organic substances in the retention index system. *Adv. Chromatogr.* 1:229–47
77. Lee ML, Vassilaros DL, White CM. 1979. Retention indices for programmed-temperature capillary-column gas chromatography of polycyclic aromatic hydrocarbons. *Anal. Chem.* 51(6):768–73
78. Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, et al. 2009. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* 81(24):10038–48
79. Babushok VI, Linstrom PJ, Reed JJ, Zenkevich IG, Brown RL, et al. 2007. Development of a database of gas chromatographic retention properties of organic compounds. *J. Chromatogr. A* 1157(1–2):414–21

80. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, et al. 2005. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21(8):1635–38
81. Stein SE, Babushok VI, Brown RL, Linstrom PJ. 2007. Estimation of Kováts retention indices using group contributions. *J. Chem. Inf. Model.* 47(3):975–80
82. Abate-Pella D, Freund DM, Ma Y, Simón-Manso Y, Hollender J, et al. 2015. Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods. *J. Chromatogr. A* 1412:43–51
83. Boswell PG, Schellenberg JR, Carr PW, Cohen JD, Hegeman AD. 2011. A study on retention "projection" as a supplementary means for compound identification by liquid chromatography-mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments. *J. Chromatogr. A* 1218(38):6732–41
84. Boswell PG, Schellenberg JR, Carr PW, Cohen JD, Hegeman AD. 2011. Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles. *J. Chromatogr. A* 1218(38):6742–49
85. Creek DJ, Jankevics A, Breitling R, Watson DG, Barrett MP, Burgess KEV. 2011. Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: improved metabolite identification by retention time prediction. *Anal. Chem.* 83(22):8703–10
86. Aalizadeh R, Nika M-C, Thomaidis NS. 2019. Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. *J. Hazard. Mater.* 363:277–85
87. Aalizadeh R, Thomaidis NS, Bletsou AA, Gago-Ferrero P. 2016. Quantitative structure-retention relationship models to support nontarget high-resolution mass spectrometric screening of emerging contaminants in environmental samples. *J. Chem. Inf. Model.* 56(7):1384–98
88. Falchi F, Bertozzi SM, Ottonello G, Ruda GF, Colombano G, et al. 2016. Kernel-based, partial least squares quantitative structure-retention relationship model for UPLC retention time prediction: a useful tool for metabolite identification. *Anal. Chem.* 88(19):9510–17
89. Baglai A, Blokland MH, Mol HGJ, Gargano AFG, van der Wal S, Schoenmakers PJ. 2018. Enhancing detectability of anabolic-steroid residues in bovine urine by actively modulated online comprehensive two-dimensional liquid chromatography–high-resolution mass spectrometry. *Anal. Chim. Acta* 1013:87–97
90. Sun W-Y, Lu Q-W, Gao H, Tong L, Li D-X, et al. 2017. Simultaneous determination of hydrophilic and lipophilic constituents in herbal medicines using directly-coupled reversed-phase and hydrophilic interaction liquid chromatography-tandem mass spectrometry. *Sci. Rep.* 7(1):7061
91. Willmann L, Erbes T, Krieger S, Trafkowski J, Rodamer M, Kammerer B. 2015. Metabolome analysis via comprehensive two-dimensional liquid chromatography: identification of modified nucleosides from RNA metabolism. *Anal. Bioanal. Chem.* 407(13):3555–66
92. Stoll DR, Shoykhet K, Petersson P, Buckenmaier S. 2017. Active solvent modulation: a valve-based approach to improve separation compatibility in two-dimensional liquid chromatography. *Anal. Chem.* 89(17):9260–67
93. Baglai A, Gargano AFG, Jordens J, Mengerink Y, Honing M, et al. 2017. Comprehensive lipidomic analysis of human plasma using multidimensional liquid- and gas-phase separations: two-dimensional liquid chromatography-mass spectrometry versus liquid chromatography-trapped-ion-mobility-mass spectrometry. *J. Chromatogr. A* 1530:90–103
94. Stoll DR, Harmes DC, Staples GO, Potter OG, Dammann CT, et al. 2018. Development of comprehensive online two-dimensional liquid chromatography/mass spectrometry using hydrophilic interaction and reversed-phase separations for rapid and deep profiling of therapeutic antibodies. *Anal. Chem.* 90(9):5923–29
95. Wang S, Li J, Shi X, Qiao L, Lu X, Xu G. 2013. A novel stop-flow two-dimensional liquid chromatography-mass spectrometry method for lipid analysis. *J. Chromatogr. A* 1321:65–72
96. SR Johnson, Lange BM. 2015. Open-access metabolomics databases for natural product research: present capabilities and future potential. *Front. Bioeng. Biotechnol.* 3:22
97. Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, et al. 2018. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* 37(4):513–32

98. Zierer J, Jackson MA, Kastenmüller G, Mangino M, Long T, et al. 2018. The fecal metabolome as a functional readout of the gut microbiome. *Nat. Genet.* 50(6):790–95
99. Noronha A, Modamio J, Jarosz Y, Guerard E, Sompairac N, et al. 2019. The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Res.* 47(D1):D614–24
100. Shaffer M, Thurimella K, Quinn K, Doenges K, Zhang X, et al. 2019. AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data. *BMC Bioinform.* 20:614
101. Fiehn O, Robertson D, Griffin J, van der Werf M, Nikolau B, et al. 2007. The metabolomics standards initiative (MSI). *Metabolomics* 3(3):175–78
102. Wallace WE, Ji W, Tchekhovskoi DV, Phinney KW, Stein SE. 2017. Mass spectral library quality assurance by inter-library comparison. *J. Am. Soc. Mass Spectrom.* 28(4):733–38
103. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. 2013. InChI—the worldwide chemical structure identifier standard. *J. Cheminform.* 5(1):7
104. Kind T, Scholz M, Fiehn O. 2009. How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLOS ONE* 4(5):e5440
105. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, et al. 2016. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 44:D463–70
106. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, et al. 2013. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41:D781–86
107. Nothias L-F, Petras D, Schmid R, Dührkop K, Rainer J, et al. 2020. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* 17(9):905–8
108. Demarque DP, Dusi RG, de Sousa FDM, Grossi SM, Silvério MRS, et al. 2020. Mass spectrometry-based metabolomics approach in the isolation of bioactive natural products. *Sci. Rep.* 10:1051