HHS Public Access

Author manuscript

Annu Rev Anal Chem (Palo Alto Calif). Author manuscript; available in PMC 2019 August 30.

Published in final edited form as:

Annu Rev Anal Chem (Palo Alto Calif). 2019 June 12; 12(1): 177–199. doi:10.1146/annurev-anchem-061318-114959.

Challenges in Identifying the Dark Molecules of Life

María Eugenia Monge[1], **James N. Dodds**[2], **Erin S. Baker**[2], **Arthur S. Edison**[3], **Facundo M. Fernández**[4]

[1]Centro de Investigaciones en Bionanociencias (CIBION), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), C1425FQD, Ciudad de Buenos Aires, Argentina

[2]Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27695, USA

[3]Department of Genetics and Biochemistry and Molecular Biology, Complex Carbohydrate Research Center, University of Georgia, Athens 30602, USA

[4]School of Chemistry and Biochemistry, Georgia Institute of Technology and Petit Institute for Biochemistry and Bioscience, Atlanta, Georgia 30332, USA

Abstract

Metabolomics is the study of the metabolome, the collection of small molecules in living organisms, cells, tissues, and biofluids. Technological advances in mass spectrometry, liquid- and gas-phase separations, nuclear magnetic resonance spectroscopy, and big data analytics have now made it possible to study metabolism at an omics or systems level. The significance of this burgeoning scientific field cannot be overstated: It impacts disciplines ranging from biomedicine to plant science. Despite these advances, the central bottleneck in metabolomics remains the identification of key metabolites that play a class-discriminant role. Because metabolites do not follow a molecular alphabet as proteins and nucleic acids do, their identification is much more time consuming, with a high failure rate. In this review, we critically discuss the state-of-the-art in metabolite identification with specific applications in metabolomics and how technologies such as mass spectrometry, ion mobility, chromatography, and nuclear magnetic resonance currently contribute to this challenging task.

Keywords

metabolomics; metabolite identification; chromatography; ion mobility; tandem mass spectrometry; nuclear magnetic resonance spectroscopy

INTRODUCTION

Metabolomics is the newest omics field focused on the examination of metabolites in complex systems, with the goal of identifying pathway alterations that correlate with the onset and progression of specific processes, such as disease (1, 2). The metabolome is

facundo.fernandez@chemistry.gatech.edu.

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.

typically defined as the collection of small molecules in a given biological system that roughly falls under the ~1,500-Da molecular weight window. Metabolites in the metabolome include endogenous molecules that are biosynthesized in primary metabolism, specialized secondary metabolite signaling molecules, lifestyle or environmental exposure molecules (the exposome), and molecules originating from the microbial community associated with the organism under study (the microbiome). Metabolomics can be either targeted to detect and quantify a set of known metabolites, or nontargeted, where the emphasis is to detect as many compounds as possible, even if associated with unknown chemical species. The metabolome, by definition, spans a vast diversity of chemistries, including lipids, sugars, amino acids, steroids, and a whole array of molecule types. These species exist in large concentration ranges that can be as high as millimolars and as low as femtomolars, suggesting that no single analytical method in existence today is able to detect, identify, and quantify all present species (3). Nuclear magnetic resonance (NMR) and liquid chromatography mass spectrometry (LC-MS) are the main platforms used for metabolomics, each with its own advantages and disadvantages in terms of sensitivity and peak capacity. These techniques are used individually in most studies but can also be combined to better yield metabolome coverage and enable more accurate metabolite identity annotation (4–6).

Despite rapid technological advances in both NMR and LC-MS, metabolite identification still remains the undisputed bottleneck in nontargeted metabolomics experiments. Only a small fraction, less than 2–10%, of the detected compounds in a nontargeted metabolomics study can be annotated reliably, with the chemical identity of most species detected remaining unknown (7). This vast chemical space has been described as the metabolome dark matter (8) and continues to be the focus of intense scientific research, in terms of new hardware and methods for improved separations and structural identification, but also new databases, libraries, and algorithms that can predict some of the analyte's molecular properties.

Much effort has been directed toward defining the confidence levels and minimum reporting standards associated with metabolite identification in any given metabolomics study. As early as 2007, Sumner et al. (9) proposed a four-level metabolite identification confidence scheme that divides metabolites into (*a*) identified compounds, (*b*) putatively annotated compounds, (*c*) putatively characterized compounds, and (*d*) unknown compounds. In level *a*, identified compounds are those that include at least two orthogonal molecular characterization methods that match a chemical standard [e.g., retention time (RT) and high-resolution mass spectrum, RT and NMR resonances, elemental formula and tandem MS (MS/MS) spectra, ^{1}H and/or ^{13}C NMR, and two-dimensional (2D) NMR spectra] (9). Generation of more than two pieces of such orthogonal metabolite characterization data is widely seen as providing good evidence in metabolite annotation, but it is not always possible due to a number of factors, including low signal-to-noise ratios, unavailability of chemical standards, and a lack of database coverage. Putatively annotated compounds are those that are identified by matching to literature or database information, but for which no chemical standard can be obtained for comparison purposes. Putatively characterized compounds are those for which only similarities to a given family of compounds can be established, but further information is not available [e.g., a glycerophospholipid characterized by the presence of an m/z 184 fragment in MS/MS experiments corresponding

to the phosphocholine headgroup; or the neutral loss of 141 Da corresponding to the phosphatidylethanolamine headgroup (10, 11)].

Despite its usefulness, this four-level scheme lacks granularity and detail, so here we propose to further expand it through a score card (Table 1) that further refines the assignment of metabolite identification confidence in nontargeted metabolomics experiments using a points system. This system builds on advances in metabolomics instrumentation and algorithms described below, such as RT prediction, use of ion mobility (IM) collision cross sections (CCS), and prediction of MS/MS fragmentation and NMR spectra.

An example of how to apply the scoring approach described in Table 1 would be as follows: If an identity is proposed for an unknown metabolite by (*a*) matching the $[M+H]^+$ and $[M+Na]^+$ monoisotopic adduct ions (5 points) with an average 2.5-ppm error (10 points), followed by (*b*) an MS^2 (10 points) and MS^3 (5 points) match to the mzCloud database, (*c*) its experimentally measured CCS matched within tolerance to a database value (10 points), and (*d*) its ^1H NMR spectrum matched to a database entry (15 points), a confidence score of (10 + 5) + (10 + 5) + (10) + (15) = 55 points could be assigned. Further improving identification confidence would require (*a*) increasing mass accuracy to 1–2 ppm for more than one adduct ion (5 points) while employing ultrahigh-resolution MS experiments to examine isotopic fine structure (20 points); (*b*) manually interpreting (5 points) the database-matched product ion spectrum to verify that observed fragment ions are compatible with known fragmentation pathway; (*c*) matching chromatographic RT to a chemical standard (10 points); (*d*) and performing 2D instead of one-dimensional (1D) NMR experiments (20 points). This would yield a more confident score of (5 + 20) + (10 + 10 + 5) + (10) + (10) + (20) = 90 points in total. By applying these more specific metabolite identification techniques to unknown species and combining the information they provide, a better confidence score is obtained. Strengths and limitations of each of these techniques, as applied to metabolomics, are discussed in the sections below.

CHROMATOGRAPHIC SEPARATIONS

Historically, chromatographic separations coupled to MS have offered one of the most versatile platforms for complex sample analysis and metabolite identification in nontargeted metabolomics studies (12). LC-MS is by far the most popular MS-based hyphenated technique in metabolomics due to its sensitivity, selectivity, reproducibility, and versatility for analyzing small molecules with a wide range of different physicochemical properties in biological samples (13, 14). Gas chromatography (GC)-MS, a predecessor to LC-MS, continues to be the obvious choice for fingerprinting volatile and low molecular weight compounds (15, 16). Analysis of less volatile compounds by GC-MS, however, requires additional preparation steps such as lyophilization, followed by chemical derivatization to increase thermal stability and volatility prior to analysis (17). To a lesser extent, capillary electrophoresis has also been utilized in metabolomics studies involving highly charged and polar ionogenic metabolites in small-volume biological samples (13, 18). Capillary electrophoresis is attractive in the sense that it separates compounds based on their intrinsic electrophoretic mobility, a parameter that reflects the charge and size of the analyte (18).

LC-MS offers the widest coverage and most efficient separation of complex metabolomes, with much enhanced peak capacity compared to MS alone (13). LC and MS are coupled by means of soft ionization techniques; electrospray ionization (ESI) (19–21) is the most common. ESI generally produces protonated or deprotonated gas-phase ions that can be used for accurate predictions of metabolite elemental formulae (22). In addition to in-source fragment ions, various adduct ions and multiply charged species that complicate data interpretation are also produced, leading to higher false-positive rates in metabolite identification (23). GC, on the other hand, is coupled to MS by means of electron ionization sources operated under vacuum and standardized at 70 eV, causing predictable fragmentation and rearrangement reactions that lead to highly reproducible mass spectra for comparison with libraries (16).

Reverse phase (RP) and hydrophilic interaction chromatography (HILIC) are by far the most commonly used chromatographic methods for LC-MS-based metabolomics (13, 24). In RPLC-MS, analytes are typically eluted by an aqueous-based mobile phase under a gradient of increasing organic solvent content from a hydrophobic stationary phase. In HILIC, analytes are eluted in order of increasing hydrophilicity from a hydrophilic stationary phase as the polarity of the mobile phase is increased by increasing the aqueous content.

Traditionally, chromatographic identification in metabolomics by either RP or HILIC approaches is best performed by RT matching with authentic chemical standards. However, not all metabolites are commercially available, and purchasing authentic standards for all possible metabolite candidates in a nontargeted study is incredibly costly and inefficient. Indeed, there are fewer reference standards available than the number of peaks that is detected in LC-MS experiments of biological samples (25–27). Although chemical synthesis can be attempted for specific cases of high-value unknowns, this degree of effort is seldom warranted.

GC-MS capillary columns are highly reproducible, which facilitates the compilation of standardized retention indices (28, 29) in libraries or public databases for compound identification (16, 30). In particular, Fiehn and collaborators (30) have greatly contributed to the metabolomics field by building FiehnLib libraries comprising mass spectra and retention indices from quadrupole and time-of-flight (ToF) GC-MS data. In contrast, it is difficult to catalog RT information from LC experiments in libraries due to the lack of procedural standardization and instrumentation involved, with varying LC pumps and injectors, columns, run temperatures, solvent gradients, mobile phase pH, and flow rates. In addition, column aging, temperature changes, MS detector drift, and other analytical factors may further compound RT variance in nontargeted LC-MS metabolomics experiments. Matrix effects caused by differences in biofluids matrix compositions can also lead to RT shifts. Spiking experiments with chemical standards can mitigate these effects (31), but in cases where chemical standards are not available, RT window prediction can complement metabolite identification efforts.

Different attempts have been reported in the literature (31–36) to integrate RT window prediction into the metabolite annotation process, most of which rely on quantitative structure–retention relationship (QSRR) modeling (37). The generalization of such

predictive models depends on the application domain investigated, because using overly restricted domains for model building also leads to poor prediction in independent test sets (38). Molecular descriptors (39–42) also play a significant role in defining such application domains (38, 43, 44), as well as in defining the multiple algorithmic options and combinations of adjustable parameters involved in QSRR model building (42, 45). In general, RT factors (RF in the equation below, where T_0 represents the chromatographic column void time) are calculated to allow comparison between different chromatographic systems (32, 36). Other studies utilize retention indices (28, 29), which are measures of relative RT based on reference compounds that elute immediately prior to and immediately following the analyte of interest (30, 45–48).

$$RF = \frac{RT - T_0}{T_0}.$$

Cao et al. (31), for example, conducted QSRR modeling based on theoretical molecular descriptors and experimental RTs of 93 authentic compounds analyzed with HILIC LC-MS. A predictive QSRR model based on a random forest algorithm achieved high predictive accuracy, with mean and median absolute errors of 0.52 min and 0.34 min (5.1% and 3.2%), respectively. These authors applied this model to annotate features (RT, *m*/*z*) of perennial ryegrass (*Lolium perenne*) samples, significantly reducing the number of false-positive metabolite annotations that would be obtained if only accurate masses were considered. Predicted RTs were validated in this study using either authentic compounds or ion fragmentation patterns (31). Among the molecular descriptors utilized to build QSRR models, the partition coefficient (XLogP) was found to be the most relevant predictor in agreement with the HILIC model previously reported by Creek et al. (32). These authors converted the RT of 120 metabolite standards to retention factors that were input into a multiple linear regression model. The optimal QSRR model used six physicochemical variables and showed good predictive ability (cross-validated $R^2 = 0.82$ and mean square error = 0.14) for RTs of metabolites with MW <400. Availability of predicted RTs translated into the removal of 40% of the false metabolite identifications based on accurate mass alone (32). The model was evaluated to putatively identify 690 metabolites in extracts of the protozoan parasite *Trypanosoma brucei*. Model limitations were associated with the applicability to only low-MW metabolites, since larger compounds were poorly predicted, most likely due to errors associated with predicted log D, the octanol-water partition coefficient calculated at a pH of 3.5. Based on their QSRR model, the authors produced a template file that allows users to calculate predicted RTs for a database of metabolites based on experimental RTs measured for standards.

QSRR modeling has also been applied to RPLC data. Bruderer et al. (34) evaluated RT prediction for a metabolomics database with 532 human metabolites. The authors built a model with only 16 compounds using logD2 (calculated based on log P and pKa) and the molecular volume as molecular descriptors. The developed model was evaluated for two different RP C18 columns and two pH conditions (pH = 3.0 and 8.0 for positive and negative ESI modes, respectively), achieving good prediction accuracy for a time window below 4 min (34). In addition, RT prediction combined with the data-independent acquisition (DIA)

method known as sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH) aided in annotating isobaric metabolites found in human urine, which increased identification confidence and reduced the number of false positives (34).

A different RT prediction model for RPLC was built by Wolfer et al. (44) based on 442 authentic standards, including fatty acids, nucleosides, sterols, sphingolipids, lipids, vitamins, cofactors, amino acids, aromatic biogenic acids, carbohydrates, catechols, neurotransmitters, and other metabolites to cover a large variety of polarity and chemical topology. The authors combined random forest and support vector regression models with 97 computationally determined descriptors derived from chemical structures. The model was further tested on an external validation set of 111 known compounds, predicting RTs with an average 13% error that reduced by 77% the number of incorrect candidates. The models were made available through a user-friendly interface to be integrated into existing workflows.

Identification of isobaric lipids is challenging even with high-resolution mass spectrometers. Aicheler et al. (33) developed an RT prediction model based on machine learning approaches that enabled the improved assignment of lipid structures and automated annotation of lipidomics data obtained by RPLC–high-resolution MS. A support vector regression model was built based on 201 lipids originating from mouse adipose tissue, including molecular structural features. The cross-validated model achieved good correlation ($R = 0.989$) between predicted and experimental RT in test samples (Figure 1) and allowed filtering out of more than half of the potential identifications, while retaining 95% of the correct candidates.

Most QSRR approaches have been successful at predicting RT with good accuracy. The main limitations of these strategies, however, are related to the lack of large and sufficiently diverse metabolite training sets for building predictive models that could cover the thousands of metabolites found in biological samples. In addition, some published models have lacked comprehensive external validation, thus risking overfitting (36). Further drawbacks are associated with the type of stationary phase material used (36, 43) that determines, together with the sample preparation protocol, the metabolome fraction that can be effectively resolved.

Alternative efforts have involved the development of tools that allow crowd sourcing of RT information across laboratories and chromatographic systems such as retention projection (49) and direct mapping (50) for GC-MS and LC-MS systems, respectively. The retention projection methodology for GC-MS data, for example, was shown to be threefold more accurate than retention indexing across five different laboratories under identical experimental conditions. This made it easier to account for unintentional differences between the various GC systems, such as temperature calibration errors, flow rate nonidealities, or variance in column dimensions (49). RT mapping for LC-MS data, developed under the name of PredRet, was done pairwise between LC systems using the same chromatographic method and provided higher accuracy than QSRR (50). However, QSRR models are able to a priori predict the RT of any given metabolite structure, whereas

PredRet can only predict RT for compounds already in the database, i.e., those for which the RT has been previously determined in a comparable chromatographic system (50).

MASS SPECTROMETRY

MS/MS experiments select a precursor ion for activation and fragmentation due to ion-neutral collisions with the gas filling the collision cell. The combination of chromatographic separations and MS/MS is one of the most powerful approaches to metabolite and natural product identification (51). In nontargeted metabolomics, these tandem mass spectra can be obtained in either a data-dependent acquisition (DDA) or DIA fashion (52). DIA approaches, such as MS^E, benefit from larger precursor ion coverage (53), but they suffer from difficulties in matching precursor ions to product ions when chromatographic overlap is substantial. DDA, on the other hand, preferentially targets the most abundant precursor ions, resulting in poorer sampling of lower abundant precursor species. Repeated analysis can somewhat improve precursor ion coverage, but this improvement is typically only marginal due to redundant precursor ion sampling. Advanced precursor ion selection algorithms used in proteomics (54) could also benefit metabolomic studies. Broeckling et al. (55), for example, have proposed an alternative method for enabling the comprehensive MS/MS coverage of complex samples via data set–dependent acquisition. With this approach, real-time feedback between data processing and data acquisition was achieved using a combination of R, ProteoWizard, XCMS, and WRENS software, yielding a threefold improvement in the number of peaks mapped by MS/MS. Elaborate DDA approaches that combine with higher-energy collision dissociation have also been reported (56).

Targeted metabolomics approaches are different from nontargeted approaches in the sense that they typically rely on chemical standards with known structures, using triple quadrupole mass spectrometers in multiple reaction monitoring (MRM) mode, or hybrid quadrupole ToF/Orbitrap analyzers in SWATH or parallel reaction monitoring modes (57). Approaches that bridge the targeted and nontargeted methods, however, have been reported in the literature. Chen et al. (58) recently developed a hybrid method utilizing DIA for targeted quantitative metabolomics experiments. In this method, a sequentially stepped targeted MS/MS scan is used for improving coverage. In this type of scan, multiple product ion scans are acquired for all ions in the examined m/z ranges, selecting them as the chosen precursor ions. These scans are then followed by scheduled MRM scans for numerous ion pairs that are used for quantitation. Ferreira et al. (59) have also reported an approach that bridges classical targeted and nontargeted methods. This approach, named MRM profiling, makes use of numerous product-precursor ion transitions that are developed in a supervised fashion. By combining neutral loss and precursor ion scans, a list of more than 1,000 transitions is built from a pooled sample. Relative abundances of all product ions targeted in these MRM transitions are then measured for all samples and used to build univariate and multivariate diagnostic models for a specific condition, such as polycystic ovarian syndrome (60).

Precursor ion coselection is an underappreciated issue that commonly hinders metabolite identification in LC-MS/MS metabolomics, particularly with lipids. Most high-resolution mass spectrometers use a low-resolution quadrupole mass analyzer for mass selection prior to MS^2 fragmentation. The selection window for the precursor ions is typically limited to

0.5–3 Da. When a high number of isobaric species chromatographically coelute, these are inevitably coselected and cofragmented, yielding a product ion spectrum that is a composite of all product ions yielded from the initial precursors in that window. One way to mitigate these interferences is to perform collision-induced dissociation (CID) experiments a posteriori from IM separations. Damen et al. (61) recently reported an ultraperformance LC (UPLC) method for the separation of closely related lipid molecular species using a stationary phase incorporating charged surface hybrid technology. The chromatographic method showed excellent RT reproducibility [intraassay relative standard deviation <0.385% and <0.451% for 20- and 10-min gradients, respectively (N = 5)]. The UPLC system was coupled to a hybrid quadrupole ToF mass spectrometer, equipped with a traveling wave ion mobility (TWIM) cell. Despite the use of a quadrupole mass analyzer for precursor ion selection, separations in the TWIM cell followed by transfer cell ion activation enabled the acquisition of cleaner low- and high-energy DIA MS/MS spectra that were more useful in terms of metabolite identification. Another approach to prevent precursor ion coselection is through the use of different variations of stored waveform inverse Fourier transform (SWIFT) ion excitation in Fourier transform ion cyclotron resonance (FT-ICR) (62). As early as 1995, O'Connor & McLafferty (63) reported on high resolution ion isolation using a capacitively coupled FT-ICR open cell. In these experiments, isotopic peaks of ubiquitin (8.6 kDa) and carbonic anhydrase (29 kDa) were isolated by SWIFT with an order of magnitude higher isolation power than previously reported in the literature. Another approach yielding high-resolution ion selection in FT-ICR MS is known as correlated harmonic excitation fields (64). With this approach, de Koning et al. (64) were able to achieve a resolution of ~50,000 in the separation of deuterated toluene isotopes. Routine implementation of high-resolution precursor ion selection approaches such as those described above in a metabolomics context could significantly improve the odds of correct unknown identification via more selective CID experiments and higher confidence MS/MS database matching.

Although MS^1 information coupled with local network enhancement analysis can be used for tentative metabolite annotation (65), high-quality tandem mass spectral libraries are becoming essential for more confident identification (66). A growing number of these libraries are currently available, including the Human Metabolome Database (HMDB) (67, 68), METLIN (69, 70), MassBank of North America (MoNA; http://mona.ftehnlab.ucdavis.edu/), LipidBlast (71), mzCloud (72), LIPID MAPS Structure Database (64, 73), Manchester Metabolomics Database (MMD) (23), and many others (74, 75). More general databases, such as PubChem (65, 76) and ChemSpider (65, 77), can also be incredibly useful for MS-based identification of unknowns. However, metabolite identification through mass spectral library searches is far from an automated task, and the analyst is typically forced to manually search each individual database and manually curate the obtained matches (if any) to ensure that differences in the type of mass spectrometer used and the collision energy employed are considered. Along these lines, Stein and coworkers (78) have proposed an approach for creating high-quality ESI tandem mass spectral libraries. The procedure involved the acquisition of tandem mass spectra for all major precursor ions in a direct infusion experiment. This was followed by assigning spectra to clusters and creating a consensus spectrum. Filtering through intensity-based constrains

for cluster membership was then applied, together with peak testing, noise reduction, and examination by an experienced human evaluator, yielding a library of >9,000 compounds with ~230,000 spectra.

When MS/MS database matches are not found, prediction of such spectra in silico can be advantageous to provide an added level of confidence to the proposed metabolite identity. To this purpose, Wolf and coworkers (79, 80) described the popular package named MetFrag, where a candidate list of metabolite identities is first obtained by searches of precursor ion masses, followed by ranking based on the agreement between in silico fragmentation spectra and experimental data. Initial evaluation of MetFrag showed that it was able to rank most of the correct compounds in the top three candidates produced by KEGG queries, producing better results than commercial software. In related work, Duhrkop et al. (81) described an approach named CFM-ID that combines computation and comparison of elemental formulae fragmentation trees with machine learning techniques. They reported a 2.5-fold increase in correct identifications compared with state-of-the-art methods when searching PubChem. Li's group (82) described a web interface (https://www.MyCompoundID.org) for metabolite identification based on both MS and MS/MS data for compounds in the HMDB and metabolites derived from these through in silico metabolic reactions. Fragmentation prediction for specific metabolite families, such as lipids, has certainly benefited subfields such as lipidomics (83–85), where chemical structures follow combinatorial rules. Prediction of electron impact spectra has also been achieved with a significant degree of success (86) based on an approach previously used for ESI data (87).

Combined prediction of various molecular properties, including retention indices, energy required to fragment 50% of a selected precursor ion, IM drift time, and CID spectrum, has been proposed by Grant and coworkers (88) through a package known as MolFind, but follow-up validation of such an approach through comparison with large experimental data sets has not been reported. Similar motivation led Hu et al. (89) to attempt the simultaneous prediction of both RTs and fragmentation patterns with the goal of identifying micropollutants.

ION MOBILITY SPECTROMETRY

In order to increase confidence in the identification of prioritized features after MS analysis, additional structural techniques, including infrared spectroscopy, NMR spectroscopy, and more recently, IM spectrometry (IMS), are often incorporated into nontargeted studies (90, 91) and are taken into account in our proposed identity confidence scoring scheme (Table 1). IMS is a gas-phase separation technique in which analytes are separated based on their rotationally averaged surface area or CCS. Briefly, IMS separations are conducted as analyte migration through an inert buffer gas under the influence of an applied electric field. Although specifics of gas composition, pressure, and applied field strength vary depending on the instrument configuration, interactions between these forces drive ion motion and separation in the IMS cell (92, 93). Because IMS distinguishes analytes based on their structural size in the gas phase, it is orthogonal to MS to a great extent, and it provides capabilities such as isomeric separations that are not possible with only the mass dimension (94, 95). Furthermore, IMS measurements occur on a millisecond timescale, which is readily

nested into traditional LC and GC-MS workflows (96). Descriptions of various IMS platforms and configurations are available in several in-depth review articles (74, 97, 98) and are beyond the scope of this article.

Although growth in MS databases has steadily continued for several decades, CCS databases are still in their infancy. Since the commercialization of IMS-MS instrumentation in 2006 (73, 76), several studies have been devoted to collecting CCS values on a larger scale. These efforts have made CCS values available to the public, while also creating databases useful for nontargeted metabolomics experiments (77, 91, 99–101). Despite these advances, obtaining reproducible CCS values across various instrument platforms still remains challenging owing to several key factors that include variations in instrument design, experimental parameters, and calibration protocols. For example, only drift tube and differential mobility instruments can empirically measure CCS from first principles, meaning that no calibrants are needed for the values they produce. IMS platforms based on TWIM spectroscopy and trapped ion mobility spectroscopy (TIMS) instruments, however, must be calibrated with ions of known mobility before reliable CCS values can be generated (102–104). Also, the specific manner in which each instrument is calibrated can greatly influence the reproducibility of the reported CCS values. Several studies have previously described the challenges of choosing the correct calibrant ions for traveling wave devices with similar structural characteristics as the species being studied (105, 106). For example, Gelb et al. (107) observed that improper calibration of TWIMS with calibrant ions of a different chemical class and charge state could produce CCS measurements with an error in excess of 4% compared to using proper calibration protocols. Even after proper calibrant ions are chosen for a specific instrument and experiment, the optimal mathematical calibration procedure remains a topic of debate for several platforms (108, 109). As IMS-MS instrumentation continues to improve in terms of generating reliable and reproducible CCS values, slight variations in calibration procedures become increasingly critical. For example, a recent interlaboratory study from Paglia et al. (77) characterized TWIMS reproducibility to typically less than 3% relative standard deviation, suggesting that errors in CCS calibration could lead to errors in calculated CCS that are larger than the instrument's own reproducibility. In a similar fashion, a recent interlaboratory study by Stow et al. (100) demonstrated that new advancements in drift tube technology resulted in reproducibility of typically less than 1% relative standard deviation for CCS measurements between laboratories because no calibration was required.

Despite challenges in calibration procedures, once a reproducible CCS value is measured for a given analyte, matching a molecular feature to a library entry is a straightforward process when the analyte m/z and CCS are compared to entries based on analytical standards. If there is a structural match within a certain mass error and CCS tolerance, the molecule is considered a match. However, if a database search generates no matches, further work is needed. CCS databases are typically generated based solely on commercially available analytical standards. Unfortunately, the availability of such standards limits the size of databases, as many compounds either cannot be isolated or they are simply too expensive to obtain. For these molecules, CCS matching may only be feasible against predicted values generated by computational methods (110–112). Such computational approaches have shown promise in generating CCS values, usually with <2% agreement with experimental

values (113). As more CCS values are published for the various IMS platforms, it is expected that predicted values will be able to fill the gap of unavailable standards for the identification of unknown metabolites.

IMS-MS also allows identification of metabolites through monitoring the placement of the analyte's observed CCS as a function of the measured m/z. For example, lipid molecules are typically characterized by their headgroup, length of the fatty acid tails, and the number of double bonds (114). These characteristics make the three-dimensional gas-phase structure of lipids quite rigid and, as a consequence, lipids have, on average, larger CCS values than peptides, carbohydrates, and nucleotides with similar masses, as shown in Figure 2 (99, 115). Several studies have noted these resulting mass/CCS ratios and have generated analytical trend lines describing the relationship between analyte mass and CCS values for a wide range of biomolecules (116, 117). In fact, if both the mass and CCS of an unknown analyte are accurately known, it may be possible to classify an unknown metabolite into a tentative biological class, after including other related factors such as mass defect and isotope ratio pattern. It is also worth noting that while there is significant overlap for several zones of the illustrated mass/CCS relationships in Figure 2, further advances in IMS resolving power and selectivity continue to increase the likelihood of producing better-resolved features in the CCS dimension.

NUCLEAR MAGNETIC RESONANCE

NMR spectroscopy exploits the quantum mechanical interactions of atomic nuclei with an external magnetic field. These interactions arise because some nuclei have an intrinsic type of angular momentum called spin. In metabolomics applications, the most common nucleus to measure is ^{1}H, with some applications using ^{13}C (e.g., 118–121), ^{31}P (e.g., 122), or other nuclei. All of these biological nuclei are spin 1/2, which means that when they are in an external magnetic field, they can adopt two energy levels separated by the resonance frequency (ω_0) that is proportional to the magnetic field strength (B_0), according to the equation

$$\omega_0 = -\gamma B_0,$$

where γ is the gyromagnetic ratio, which is a physical constant for a specific nucleus. Thus, ^{1}H resonates at 600 MHz in a 14.1-Tesla (T) magnet and 900 MHz in a 21.1-T magnet. Because of the quantum mechanical underpinnings of NMR spectroscopy, it provides atom-specific information, which makes it the method of choice for the structural characterization of unknown molecules. It can be quantitative, with the integrated value of each NMR proportional to the number of nuclei and thus the concentration of the molecule. NMR spectroscopy is also nondestructive and highly reproducible because the sample never comes into direct contact with the instrument. But all of these significant strengths come at a cost of overall sensitivity. Because it is a resonance phenomenon, NMR spectroscopy has a fundamental sensitivity that is limited by the Boltzmann equation

$$\frac{N_{\text{up}}}{N_{\text{down}}} = e^{-\left(\frac{\Delta E}{k_B T}\right)},$$

where N_{up} and N_{down} represent the number of nuclear spins in the upper and lower energy levels, ΔE is the energy gap between levels, k_B is the Boltzmann constant, and T is the absolute temperature. For ^{1}H at 600 MHz and room temperature, if the number of spins in the upper energy state is 1 million, there are only 1 million + 96 spins in the lower state, so only a small fraction of the sample contributes to the NMR signal. But the low energies associated with NMR spectroscopy also allow its noninvasive application in living systems through magnetic resonance imaging.

ONE-DIMENSIONAL NUCLEAR MAGNETIC RESONANCE

NMR metabolomics applications are typically done without chromatography or significant sample extraction steps. Therefore, the measured signals represent a complex mixture of metabolites in a sample. With modern spectrometers and probes, the practical lower limit of detection is about 10 μM in a 550-μL sample. The most common experiment in NMR metabolomics is a 1D ^{1}H spectrum, which can have hundreds to sometimes thousands of overlapping peaks. These can be matched to databases with standard spectra of known metabolites. The most important public databases with NMR spectral libraries are the Biological Magnetic Resonance Data Bank (BMRB) (123) and HMDB (124). The primary difficulty in using these databases is that 1D NMR spectra can be heavily overlapped and, thus, there are almost always uncertainties in peak assignment using exclusively 1D methods.

TWO-DIMENSIONAL NUCLEAR MAGNETIC RESONANCE

Two-dimensional NMR offers significant advantages over 1D NMR. It not only reduces resonance overlap by spreading the signal into a second dimension, but it also can provide extra information about chemical bonding between nuclei. The drawback of 2D NMR is the length of time required for each experiment, so these are typically only used for pooled samples rather than every sample in a study. However, new approaches are improving the speed of 2D methods (125). One of the most useful 2D experiments in metabolomics is heteronuclear single quantum correlation (HSQC). The 2D HSQC experiment correlates ^{1}H with ^{13}C (or less common in metabolomics, ^{15}N) that are covalently bonded. Each pair of bonded nuclei give a single peak in a 2D HSQC, and this provides a useful fingerprint of a mixture. Edison & Schroeder (126) wrote a more complete description of 2D NMR experiments and their interpretation.

NUCLEAR MAGNETIC RESONANCE DEREPLICATION

Before the difficult step of unknown compound identification, it is important to first recognize and assign peaks that are known and in databases. This is called dereplication. There are several approaches to this, as both freely available and commercial packages. One

of the most popular commercial software is Chenomx, which allows users to fit a library of reference standards to 1D ^{1}H experimental metabolomics data. This is an excellent visualization tool but suffers from the problems mentioned above about 1D NMR and peak overlap. At least one study reports inconsistent results with Chenomx using the same data set and multiple analysts (127). Bruker Corporation offers a spectral database that includes a wide range of pH values for many common metabolites, and this can be used with both 1D and 2D data in its Assure software. The 2D data add confidence in annotation, but the cost of this solution may be beyond the budget of many labs.

Brüschweiler's laboratory has developed a suite of free web-based tools called COLMAR (complex mixture analysis by NMR; http://spin.ccic.ohio-state.edu/index.php/colmar). COLMAR has several functionalities that are useful for metabolomics. One of the simplest ways to use COLMAR is the ^{1}H-^{13}C-HSQC query, which takes a peak list from an HSQC spectrum and finds database matches using a combination of BMRB, HMDB, and internally curated data. HSQC matches are useful but can also be prone to misinterpretation, because they only report on a ^{1}H–^{13}C pair and do not include correlations between peaks. COLMAR adds to the HSQC query by allowing the addition of TOCSY (total correlation spectroscopy) or HSQC-TOCSY. The TOCSY experiments provide correlations between coupled ^{1}H spins. Adding both HSQC and TOCSY or HSQC-TOCSY data significantly improves the confidence of dereplication of known metabolites in a mixture.

COMPOUND ISOLATION

The most straightforward approach to NMR-based compound identification is to use a natural products-like strategy that involves purification of the unknown molecule. The purification steps are typically activity guided, essentially using the desired activity as a detector for the compound of interest. For example, the identification of the mating pheromone for *Caenorhabditis elegans* involved a series of low-resolution fractionations followed by assays of male-specific attraction (128). Once a fraction is sufficiently pure, both 2D NMR (and MS) data can be obtained and analyzed. There are several advantages of this strategy. First, the focus is on the compound of interest (i.e., the one with the desired biological activity or discriminating power). Second, the limits of detection are defined by the assay and not the NMR spectrometer. This is very important, because unknown molecules at concentrations lower than NMR detection limits can be concentrated and identified if sufficient material is available for the bioassay (129). Finally, the pure (or semipure) compound provides a straightforward way to relate NMR and MS data, which is important for a more reliable identification (see Table 1).

Although it is common to use some type of LC system for fractionation, it is not always best to start with analytical chromatography. It is often simpler to start with simpler solid-phase extraction (SPE) steps, which can be done on a larger scale than LC. Orthogonal SPE (e.g., C18 followed by ion exchange) can be quite effective at quickly simplifying mixtures, even with just a few fractions from each step. If necessary, the crude material from SPE fractionation can then be purified further using LC. This approach was used for the isolation of *C. elegans* (128) and *Panagrellus redivivus* (130) mating pheromones (and many other activity-guided fractionation studies).

Alternatively, the assay can be NMR or MS spectra in order to isolate an unknown peak of interest (131). In most metabolomics applications, important NMR or MS features are first determined from statistical analysis. If these features do not match databases, the same purification steps described for activity-guided fraction can be used, and fractions can be screened for the feature(s) of interest. Chemical fractionation is time consuming and not always possible without sample degradation (132).

DIFFERENTIAL ANALYSIS BY TWO-DIMENSIONAL NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

Frank Schroeder's lab has developed a technique called differential analysis by 2D NMR spectroscopy (DANS) (133). A review of DANS and other approaches to NMR mixture analysis provides an additional overview (134). Briefly, DANS compares unfractionated high-resolution 2D NMR data sets (COSY) of two different genetic strains of organism, e.g., wild-type and a mutant of interest. In DANS, the data are manually overlaid, subtracted, and adjusted so that peaks common to both spectra will cancel, while peaks unique to one of the genotypes are retained. The DANS strategy does not include integration of MS data, but it does provide a very helpful overview of the major metabolic differences between the two genotypes and also then yields 2D NMR data that can be used to either partially or fully determine the structure of unknowns.

THE SUMMIT APPROACH

The Brüschweiler lab recently developed a powerful approach to link NMR and high-resolution MS data in metabolomics called structures of unknown metabolomic mixture components by MS/NMR (SUMMIT) (135). SUMMIT links MS and NMR data through computation (Figure 3). Chemical purification can be used but is not a requirement. The general idea is that both high-resolution MS and NMR data are collected on the same sample. The high-resolution MS can be obtained through either chromatography or direct infusion, e.g., with an FT-ICR instrument (136). Starting with a feature of interest from the MS data, it is possible to obtain a molecular formula directly from the intact precursor ion provided there is sufficiently high mass resolution.

Once a reliable molecular formula is known, it is possible to enumerate all structures that are consistent with that formula, e.g., by searching the ChemSpider database (http://www.chemspider.com). The difficulty with this step is that the number of possible molecules grows substantially with molecular weight. For example, $C_4H_6O_5$ (e.g., maleic acid) yields 35 structures from ChemSpider. In contrast, ChemSpider yields 1,023 results for $C_{21}H_{30}O_3$. Therefore, it is desirable for the number of possible structures to be reduced by other data, e.g., through association with a specific metabolic pathway through a metabolite-genome-wide association study.

The next step is to calculate the NMR chemical shifts of all possible structures from MS data. Calculations of NMR chemical shifts have become quite reliable with high-level ab initio or density functional quantum mechanical calculations (137). However, even semiempirical-based methods can provide reasonable results (135). The computed NMR

chemical shifts are then compared against experimental NMR data for the closest match. This conceptually simple step can be complicated when the NMR data are from a complex and unfractionated metabolomics mixture, so a modification of this strategy would be to fractionate the NMR sample using similar chromatography to the LC-MS. By doing this step, the overall approach becomes similar to natural products fractionation described above with the additional step of using computational chemistry to determine the best structure rather than through traditional analysis.

CONCLUSIONS AND OUTLOOK

Despite the limitations associated with each of the strategies described here, all of them have significantly contributed to addressing the most challenging problem in nontargeted metabolomics studies, which is to know the unknowns (27, 69, 138). Integration of the information produced by such advanced assays, however, is still largely lacking, thus preventing identification of metabolites in a high-throughput fashion. Expected advances in metabolomics informatics pipelines are expected to propel the field to a more mature stage, in a similar fashion to what has occurred in other omics fields such as genomics, transcriptomics and proteomics.

ACKNOWLEDGMENTS

A.S.E. and F.M.F. were supported by the US National Institutes of Health (NIH) (grant 1U2CES030167–01) and the CMaT NSF Research Center (grant EEC-1648035). A.S.E. was also supported by the Georgia Research Alliance and acknowledges productive collaborations with Rafael Brüschweiler and Frank Schroeder. J.N.D. and E.S.B. would like to acknowledge support from the National Institute of Environmental Health Sciences of the NIH (P42 ES027704). F.M.F. acknowledges support from 1R01CA218664–01, a Cystic Fibrosis Foundation Research Development Program grant, and the NIH Molecular Transducers of Physical Activity Consortium (grant 1U24DK112341–01). M.E.M. is a research staff member from the CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina). M.E.M. acknowledges support from the National Agency of Scientific and Technological Promotion (PRH-PICT-2015–0022 project) and from CONICET (grant PUE 055).

LITERATURE CITED

1. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, et al. 2011 Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat. Protoc 6:1060–83 [PubMed: 21720319]
2. Vermeersch KA, Styczynski MP. 2013 Applications of metabolomics in cancer research. J. Carcinog 12:9 [PubMed: 23858297]
3. Weckwerth W, Morgenthal K. 2005 Metabolomics: from pattern recognition to biological interpretation. Drug. Discov. Today 10:1551–58 [PubMed: 16257378]
4. Hao J, Liebeke M, Sommer U, Viant MR, Bundy JG, Ebbels TM. 2016 Statistical correlations between NMR spectroscopy and direct infusion FT-ICR mass spectrometry aid annotation of unknowns in metabolomics. Anal. Chem 88:2583–89 [PubMed: 26824414]
5. Geier FM, Want EJ, Leroi AM, Bundy JG. 2011 Cross-platform comparison of *Caenorhabditis elegans* tissue extraction strategies for comprehensive metabolome coverage. Anal. Chem 83:3730–36 [PubMed: 21480661]
6. Nevedomskaya E, Mayboroda OA, Deelder AM. 2011 Cross-platform analysis of longitudinal data in metabolomics. Mol. Biosyst 7:3214–22 [PubMed: 21947311]
7. Anonymous. 2008 Metabolomics: dark matter. Nature 455:698 10.1038/455698a [PubMed: 18833282]
8. da Silva RR, Dorrestein PC, Quinn RA. 2015 Illuminating the dark matter in metabolomics. PNAS 112:12549–50 [PubMed: 26430243]

9. 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 3:211–21 [PubMed: 24039616]
10. Ding J, Sorensen CM, Jaitly N, Jiang H, Orton DJ, et al. 2008 Application of the accurate mass and time tag approach in studies of the human blood lipidome. J. Chromatogr. B Anal. Technol. Biomed. Life Sci 871:243–52
11. Koelmel JP, Ulmer CZ, Jones CM, Yost RA, Bowden JA. 2017 Common cases of improper lipid annotation using high-resolution tandem mass spectrometry data and corresponding limitations in biological interpretation. Biochim. Biophys. Acta 1862:766–70
12. Lei Z, Huhman DV, Sumner LW. 2011 Mass spectrometry strategies in metabolomics. J. Biol. Chem 286:25435–42 [PubMed: 21632543]
13. Kuehnbaum NL, Britz-McKibbin P. 2013 New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. Chem. Rev 113:2437–68 [PubMed: 23506082]
14. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, et al. 2011 Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat. Protoc 6:1060–83 [PubMed: 21720319]
15. Fiehn O 2008 Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. Trends Anal. Chem 27:261–69
16. Fiehn O 2016 Metabolomics by gas chromatography-mass spectrometry: the combination of targeted and untargeted profiling. Curr. Protoc. Mol. Biol 114:304.1–.32 [PubMed: 27038389]
17. Haggarty J, Burgess KE. 2017 Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. Curr. Opin. Biotechnol 43:77–85 [PubMed: 27771607]
18. Zhang W, Hankemeier T, Ramautar R. 2017 Next-generation capillary electrophoresis-mass spectrometry approaches in metabolomics. Curr. Opin. Biotechnol 43:1–7 [PubMed: 27455398]
19. Fenn J, Mann M, Meng C, Wong S, Whitehouse C. 1989 Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64–71 [PubMed: 2675315]
20. Kebarle P, Verkerk UH. 2009 Electrospray: From ions in solution to ions in the gas phase, what we know now. Mass Spectrom. Rev 28:898–917 [PubMed: 19551695]
21. Kebarle P 2000 A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. J. Mass Spectrom 35:804–17 [PubMed: 10934434]
22. Pluskal T, Uehara T, Yanagida M. 2012 Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. Anal. Chem 84:4396–403 [PubMed: 22497521]
23. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, et al. 2009 Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. Analyst 134:1322–32 [PubMed: 19562197]
24. Spagou K, Tsoukali H, Raikos N, Gika H, Wilson ID, Theodoridis G. 2010 Hydrophilic interaction chromatography coupled to MS for metabonomic/metabolomic studies. J. Sep. Sci 33:716–27 [PubMed: 20187037]
25. Dunn W, Erban A, Weber RM, Creek D, Brown M, et al. 2013 Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics 9:44–66
26. Kind T, Fiehn O. 2010 Advances in structure elucidation of small molecules using mass spectrometry. Bioanal. Rev 2:23–60 [PubMed: 21289855]
27. Wishart DS. 2011 Advances in metabolite identification. Bioanalysis 3:1769–82 [PubMed: 21827274]
28. Babushok V, Linstrom P. 2004 On the Relationship between Kováts and Lee retention indices. Chromatographia 60:725–28
29. Rostad CE, Pereira WE. 1986 Kovats and lee retention indices determined by gas chromatography/mass spectrometry for organic compounds of environmental interest. J. High Resolut. Chromatogr 9:328–34
30. 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 81:10038–48 [PubMed: 19928838]

31. Cao M, Fraser K, Huege J, Featonby T, Rasmussen S, Jones C. 2015 Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. Metabolomics 11:696–706 [PubMed: 25972771]
32. Creek DJ, Jankevics A, Breitling R, Watson DG, Barrett MP, Burgess KE. 2011 Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: improved metabolite identification by retention time prediction. Anal. Chem 83:8703–10 [PubMed: 21928819]
33. Aicheler F, Li J, Hoene M, Lehmann R, Xu G, Kohlbacher O. 2015 Retention time prediction improves identification in nontargeted lipidomics approaches. Anal. Chem 87:7698–704 [PubMed: 26145158]
34. Bruderer T, Varesio E, Hopfgartner G. 2017 The use of LC predicted retention times to extend metabolites identification with SWATH data acquisition. J. Chromatogr. B Anal. Technol. Biomed. Life Sci 1071:3–10
35. Hagiwara T, Saito S, Ujiie Y, Imai K, Kakuta M, et al. 2010 HPLC retention time prediction for metabolome analysis. Bioinformation 5:255–58 [PubMed: 21364827]
36. Navarro-Reig M, Ortiz-Villanueva E, Tauler R, Jaumot J. 2017 Modelling of hydrophilic interaction liquid chromatography stationary phases using chemometric approaches. Metabolites 7:54
37. Kaliszan R 2007 QSRR: quantitative structure-(chromatographic) retention relationships. Chem. Rev 107:3212–46 [PubMed: 17595149]
38. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. 2012 Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17:4791–810 [PubMed: 22534664]
39. Mauri ACV, Pavan M, Todeschini R. 2006 Dragon software: an easy approach to molecular descriptor calculations. Match 56:237–48
40. Hong H, Xie Q, Ge W, Qian F, Fang H, et al. 2008 Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J. Chem. Inf. Model 48:1337–44 [PubMed: 18564836]
41. Yap CW. 2011 PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem 32:1466–74 [PubMed: 21425294]
42. 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 88:9510–17 [PubMed: 27583774]
43. Gorynski K, Bojko B, Nowaczyk A, Bucinski A, Pawliszyn J, Kaliszan R. 2013 Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds. Anal. Chim. Acta 797:13–19 [PubMed: 24050665]
44. Wolfer AM, Lozano S, Umbdenstock T, Croixmarie V, Arrault A, Vayer P. 2015 UPLC–MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling. Metabolomics 12:8
45. Hall LM, Hill DW, Menikarachchi LC, Chen M-H, Hall LH, Grant DF. 2015 Optimizing artificial neural network models for metabolomics and systems biology: an example using HPLC retention index data. Bioanalysis 7:939–55 [PubMed: 25966007]
46. Hall LM, Hall LH, Kertesz TM, Hill DW, Sharp TR, et al. 2012 Development of Ecom50 and retention index models for nontargeted metabolomics: identification of 1,3-dicyclohexylurea in human serum by HPLC/mass spectrometry. J. Chem. Inf. Model 52:1222–37 [PubMed: 22489687]
47. Hall LM, Hill DW, Bugden K, Cawley S, Hall LH, et al. 2018 Development of a reverse phase HPLC retention index model for nontargeted metabolomics using synthetic compounds. J. Chem. Inf. Model 58:591–604 [PubMed: 29489351]
48. Stein SE, Babushok VI, Brown RL, Linstrom PJ. 2007 Estimation of Kovats retention indices using group contributions. J. Chem. Inf. Model 47:975–80 [PubMed: 17367127]
49. Barnes BB, Wilson MB, Carr PW, Vitha MF, Broeckling CD, et al. 2013 “Retention projection” enables reliable use of shared gas chromatographic retention data across laboratories, instruments, and methods. Anal. Chem 85:11650–57 [PubMed: 24205931]

50. Stanstrup J, Neumann S, Vrhovsek U. 2015 PredRet: prediction of retention time by direct mapping between multiple chromatographic systems. Anal. Chem 87:9421–28 [PubMed: 26289378]
51. Johnson AR, Carlson EE. 2015 Collision-induced dissociation mass spectrometry: a powerful tool for natural product structure elucidation. Anal. Chem 87:10668–78 [PubMed: 26132379]
52. Doerr A 2015 DIA mass spectrometry. Nat. Methods 12:35–35
53. Bateman KP, Castro-Perez J, Wrona M, Shockcor JP, Yu K, et al. 2007 MS^E with mass defect filtering for in vitro and in vivo metabolite identification. Rapid Commun. Mass Spectrom 21:1485–96 [PubMed: 17394128]
54. Kreimer S, Belov ME, Danielson WF, Levitsky LI, Gorshkov MV, et al. 2016 Advanced precursor ion selection algorithms for increased depth of bottom-up proteomic profiling. J. Proteome Res 15:3563–73 [PubMed: 27569903]
55. Broeckling CD, Hoyes E, Richardson K, Brown JM, Prenni JE. 2018 Comprehensive tandem-mass-spectrometry coverage of complex samples enabled by data-set-dependent acquisition. Anal. Chem 90:8020–27 [PubMed: 29846054]
56. Mullard G, Allwood JW, Weber R, Brown M, Begley P, et al. 2015 A new strategy for MS/MS data acquisition applying multiple data dependent experiments on Orbitrap mass spectrometers in non-targeted metabolomic applications. Metabolomics 11:1068–80
57. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. 2012 Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Mol. Cell. Proteom 11:1475–88
58. Chen YH, Zhou Z, Yang W, Bi N, Xu J, et al. 2017 Development of a data-independent targeted metabolomics method for relative quantification using liquid chromatography coupled with tandem mass spectrometry. Anal. Chem 89:6954–62 [PubMed: 28574715]
59. Ferreira CR, Yannell KE, Mollenhauer B, Espy RD, Cordeiro FB, et al. 2016 Chemical profiling of cerebrospinal fluid by multiple reaction monitoring mass spectrometry. Analyst 141:5252–55 [PubMed: 27517482]
60. Cordeiro FB, Ferreira CR, Sobreira TJP, Yannell KE, Jarmusch AK, et al. 2017 Multiple reaction monitoring (MRM)-profiling for biomarker discovery applied to human polycystic ovarian syndrome. Rapid Commun. Mass Spectrom 31:1462–70 [PubMed: 28656689]
61. Damen CWN, Isaac G, Langridge J, Hankemeier T, Vreeken RJ. 2014 Enhanced lipid isomer separation in human plasma using reversed-phase UPLC with ion-mobility/high-resolution MS detection. J. Lipid Res 55:1772–83 [PubMed: 24891331]
62. Guan SH, Marshall AG. 1996 Stored waveform inverse Fourier transform (SWIFT) ion excitation in trapped-ion mass spectometry: theory and applications. Int. J. Mass Spectrom 157:5–37
63. O'Connor PB, McLafferty FW. 1995 High-resolution ion isolation with the ion-cyclotron resonance capacitively coupled open cell. J. Am. Soc. Mass Spectrom 6:533–35 [PubMed: 24214309]
64. de Koning LJ, Nibbering NMM, van Orden SL, Laukien FH. 1997 Mass selection of ions in a Fourier transform ion cyclotron resonance trap using correlated harmonic excitation fields (CHEF). Int. J. Mass Spectrom 165:209–19
65. Li SZ, Park Y, Duraisingham S, Strobel FH, Khan N, et al. 2013 Predicting network activity from high throughput metabolomics. PLOS Comp. Biol 9:e1003123
66. Kind T, Tsugawa H, Cajka T, Ma Y, Lai ZJ, et al. 2018 Identification of small molecules using accurate mass MS/MS search. Mass Spectrom. Rev 37:513–32 [PubMed: 28436590]
67. Mandal R, Chamot D, Wishart DS. 2018 The role of the Human Metabolome Database in inborn errors of metabolism. J. Inherit. Metab. Dis 41:329–36 [PubMed: 29663269]
68. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, et al. 2018 HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46:D608–17 [PubMed: 29140435]
69. 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 90:3156–64 [PubMed: 29381867]
70. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, et al. 2005 METLIN: a metabolite mass spectral database. Ther. Drug Monit 27:747–51 [PubMed: 16404815]

71. Kind T, Liu KH, Lee DY, DeFelice B, Meissen JK, Fiehn O. 2013 LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat. Methods 10:755–58 [PubMed: 23817071]
72. Mistrik R 2018 mzCLOUD: A spectral tree library for the Identification of “unknown unknowns.” Abstr. Pap. Am. Chem. Soc 2018:255
73. Pringle SD, Giles K, Wildgoose JL, Williams JP, Slade SE, et al. 2007 An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. Int. J. Mass Spectrom 261:1–12
74. Kanu AB, Dwivedi P, Tam M, Matz L, Hill HH. 2008 Ion mobility–mass spectrometry. J. Mass Spectrom 43:1–22 [PubMed: 18200615]
75. Hu QZ, Cooks RG, Noll RJ. 2007 Phase-enhanced selective ion ejection in an Orbitrap mass spectrometer. J. Am. Soc. Mass Spectrom 18:980–83 [PubMed: 17382556]
76. Ausloos P, Clifton CL, Lias SG, Mikaya AI, Stein SE, et al. 1999 The critical evaluation of a comprehensive mass spectral library. J. Am. Soc. Mass Spectrom 10:287–99 [PubMed: 10197350]
77. Paglia G, Angel P, Williams JP, Richardson K, Olivos HJ, et al. 2015 Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. Anal. Chem 87:1137–44 [PubMed: 25495617]
78. Yang XY, Neta P, Stein SE. 2014 Quality control for building libraries from electrospray ionization tandem mass spectra. Anal. Chem 86:6393–400 [PubMed: 24896981]
79. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. 2010 In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinf 11:148
80. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. 2016 MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation. J. Cheminform 8:3 [PubMed: 26834843]
81. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. 2015 Searching molecular structure databases with tandem mass spectra using CSI:FingerID. PNAS 112:12580–85 [PubMed: 26392543]
82. Huan T, Tang CQ, Li RH, Shi Y, Lin GH, Li L. 2015 MyCompoundID MS/MS search: metabolite identification using a library of predicted fragment-ion-spectra of 383,830 possible human metabolites. Anal. Chem 87:10619–26 [PubMed: 26415007]
83. Witting M, Ruttkies C, Neumann S, Schmitt-Kopplin P. 2017 LipidFrag: improving reliability of in silico fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome. PLOS ONE 12:e0172311 [PubMed: 28278196]
84. Houjou T, Yamatani K, Imagawa M, Shimizu T, Taguchi R. 2005 A shotgun tandem mass spectrometric analysis of phospholipids with normal-phase and/or reverse-phase liquid chromatography/electrospray ionization mass spectrometry. Rapid Commun. Mass Spectrom 19:654–66 [PubMed: 15700236]
85. Kyle JE, Crowell KL, Casey CP, Fujimoto GM, Kim S, et al. 2017 LIQUID: an-open source software for identifying lipids in LC-MS/MS-based lipidomics data. Bioinformatics 33:1744–46 [PubMed: 28158427]
86. Allen F, Pon A, Greiner R, Wishart D. 2016 Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. Anal. Chem 88:7689–97 [PubMed: 27381172]
87. Allen F, Greiner R, Wishart D. 2015 Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics 11:98–110
88. Menikarachchi LC, Cawley S, Hill DW, Hall LM, Hall L, et al. 2012 MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures. Anal. Chem 84:9388–94 [PubMed: 23039714]
89. Hu M, Muller E, Schymanski EL, Ruttkies C, Schulze T, et al. 2018 Performance of combined fragmentation and retention prediction for the identification of organic micropollutants by LC-HRMS. Anal. Bioanal. Chem 410:1931–41 [PubMed: 29380019]
90. Livanos AE, Greiner TU, Vangay P, Pathmasiri W, Stewart D, et al. 2016 Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. Nat. Microbiol 1:16140 [PubMed: 27782139]
91. Mairinger T, Causon TJ, Hann S. 2018 The potential of ion mobility–mass spectrometry for nontargeted metabolomics. Curr. Opin. Chem. Biol 42:9–15 [PubMed: 29107931]

92. Dodds JN, May JC, McLean JA. 2017 Correlating resolving power, resolution, and collision cross section: unifying cross-platform assessment of separation efficiency in ion mobility spectrometry. Anal. Chem 89:12176–84 [PubMed: 29039942]

93. Harper B, Neumann EK, Stow SM, May JC, McLean JA, Solouki T. 2016 Determination of ion mobility collision cross sections for unresolved isomeric mixtures using tandem mass spectrometry and chemometric deconvolution. Anal. Chim. Acta 939:64–72 [PubMed: 27639144]

94. Groessl M, Graf S, Knochenmuss R. 2015 High resolution ion mobility-mass spectrometry for separation and identification of isomeric lipids. Analyst 140:6904–11 [PubMed: 26312258]

95. Hofmann J, Hahm HS, Seeberger PH, Pagel K. 2015 Identification of carbohydrate anomers using ion mobility–mass spectrometry. Nature 526:241–44 [PubMed: 26416727]

96. May JC, McLean JA. 2015 Ion mobility-mass spectrometry: time-dispersive instrumentation. Anal. Chem 87:1422–36 [PubMed: 25526595]

97. Cumeras R, Figueras E, Davis CE, Baumbach JI, Gràcia I. 2015 Review on ion mobility spectrometry. Part 1: current instrumentation. Analyst 140:1376–90 [PubMed: 25465076]

98. Cumeras R, Figueras E, Davis CE, Baumbach JI, Gràcia I. 2015 Review on ion mobility spectrometry. Part 2: hyphenated methods and effects of experimental parameters. Analyst 140:1391–410 [PubMed: 25465248]

99. Hines KM, Ross DH, Davidson KL, Bush MF, Xu L. 2017 Large-scale structural characterization of drug and drug-like compounds by high-throughput ion mobility-mass spectrometry. Anal. Chem 89:9023–30 [PubMed: 28764324]

100. Stow SM, Causon TJ, Zheng X, Kurulugama RT, Mairinger T, et al. 2017 An interlaboratory evaluation of drift tube ion mobility–mass spectrometry collision cross section measurements. Anal. Chem 89:9048–55 [PubMed: 28763190]

101. Bush MF, Hall Z, Giles K, Hoyes J, Robinson CV, Ruotolo BT. 2010 Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. Anal. Chem 82:9557–65 [PubMed: 20979392]

102. Michelmann K, Silveira JA, Ridgeway ME, Park MA. 2015 Fundamentals of trapped ion mobility spectrometry. J. Am. Soc. Mass Spectrom 26:14–24 [PubMed: 25331153]

103. Silveira JA, Ridgeway ME, Park MA. 2014 High resolution trapped ion mobility spectrometery of peptides. Anal. Chem 86:5624–27 [PubMed: 24862843]

104. Giles K, Williams JP, Campuzano I. 2011 Enhancements in travelling wave ion mobility resolution. Rapid Commun. Mass Spectrom 25:1559–66 [PubMed: 21594930]

105. Forsythe JG, Petrov AS, Walker CA, Allen SJ, Pellissier JS, et al. 2015 Collision cross section calibrants for negative ion mode traveling wave ion mobility-mass spectrometry. Analyst 140:6853–61 [PubMed: 26148962]

106. Hines KM, May JC, McLean JA, Xu L. 2016 Evaluation of collision cross section calibrants for structural analysis of lipids by traveling wave ion mobility-mass spectrometry. Anal. Chem 88:7329–36 [PubMed: 27321977]

107. Gelb AS, Jarratt RE, Huang Y, Dodds ED. 2014 A study of calibrant selection in measurement of carbohydrate and peptide ion-neutral collision cross sections by traveling wave ion mobility spectrometry. Anal. Chem 86:11396–402 [PubMed: 25329513]

108. Chai M, Young MN, Liu FC, Bleiholder C. 2018 A transferable, sample-independent calibration procedure for trapped ion mobility spectrometry (TIMS). Anal. Chem 90:9040–47 [PubMed: 29975506]

109. Bush MF, Campuzano IDG, Robinson CV. 2012 Ion mobility mass spectrometry of peptide ions: effects of drift gas and calibration strategies. Anal. Chem 84:7124–30 [PubMed: 22845859]

110. Zhou Z, Shen X, Tu J, Zhu Z-J. 2016 Large-scale prediction of collision cross-section values for metabolites in ion mobility-mass spectrometry. Anal. Chem 88:11084–91 [PubMed: 27768289]

111. Zhou Z, Xiong X, Zhu Z-J. 2017 MetCCS predictor: a web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. Bioinformatics 33:2235–37 [PubMed: 28334295]

112. Soper-Hopper MT, Petrov AS, Howard JN, Yu SS, Forsythe JG, et al. 2017 Collision cross section predictions using 2-dimensional molecular descriptors. Chem. Commun 53:7624–27

113. Campuzano I, Bush MF, Robinson CV, Beaumont C, Richardson K, et al. 2012 Structural characterization of drug-like compounds by ion mobility mass spectrometry: comparison of theoretical and experimentally derived nitrogen collision cross sections. Anal. Chem 84:1026–33 [PubMed: 22141445]
114. Kyle JE, Zhang X, Weitz KK, Monroe ME, Ibrahim YM, et al. 2016 Uncovering biologically significant lipid isomers with liquid chromatography, ion mobility spectrometry and mass spectrometry. Analyst 141:1649–59 [PubMed: 26734689]
115. Kliman M, May JC, McLean JA. 2011 Lipid analysis and lipidomics by structurally selective ion mobility-mass spectrometry. Biochim. Biophys. Acta 1811:935–45 [PubMed: 21708282]
116. Goodwin CR, Fenn LS, Derewacz DK, Bachmann BO, McLean JA. 2012 Structural mass spectrometry: rapid methods for separation and analysis of peptide natural products. J. Nat. Prod 75:48–53 [PubMed: 22216918]
117. May JC, Goodwin CR, Lareau NM, Leaptrot KL, Morris CB, et al. 2014 Conformational ordering of biomolecules in the gas phase: nitrogen collision cross sections measured on a prototype high resolution drift tube ion mobility-mass spectrometer. Anal. Chem 86:2107–16 [PubMed: 24446877]
118. Clendinen CS, Stupp GS, Wang B, Garrett TJ, Edison AS. 2016 ^{13}C metabolomics: NMR and IROA for unknown identification. Curr. Metab 4:116–20
119. Clendinen CS, Stupp GS, Ajredini R, Lee-McMullen B, Beecher C, Edison AS. 2015 An overview of methods using ^{13}C for improved compound identification in metabolomics and natural products. Front. Plant Sci 6:611 [PubMed: 26379677]
120. Clendinen CS, Pasquel C, Ajredini R, Edison AS. 2015 ^{13}C NMR metabolomics: INADEQUATE network analysis. Anal. Chem 87:5698–706 [PubMed: 25932900]
121. Clendinen CS, Lee-McMullen B, Williams CM, Stupp GS, Vandenborne K, et al. 2014 ^{13}C NMR metabolomics: applications at natural abundance. Anal. Chem 86:9242–50 [PubMed: 25140385]
122. Nemutlu E, Juranic N, Zhang S, Ward LE, Dutta T, et al. 2012 Electron spray ionization mass spectrometry and 2D ^{31}P NMR for monitoring ^{18}O/^{16}O isotope exchange and turnover rates of metabolic oligophosphates. Anal. Bioanal. Chem 403:697–706 [PubMed: 22427058]
123. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, et al. 2008 BioMagResBank. Nucleic Acids Res 36:D402–8 [PubMed: 17984079]
124. Wishart DS, Knox C, Guo AC, Eisner R, Young N, et al. 2009 HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37:D603–10 [PubMed: 18953024]
125. Giraudeau P, Frydman L. 2014 Ultrafast 2D NMR: an emerging tool in analytical spectroscopy. Annu. Rev. Anal. Chem 7:129–61
126. Edison AS, Schroeder FC. 2010 NMR spectroscopy of small molecules and analysis of complex mixtures. In Comprehensive Natural Products II. Chemistry and Biology, ed. Mander L, Hung-Wen L, pp. 169–96. Oxford, UK: Elsevier
127. Tredwell GD, Behrends V, Geier FM, Liebeke M, Bundy JG. 2011 Between-person comparison of metabolite fitting for NMR-based quantitative metabolomics. Anal. Chem 83:8683–87 [PubMed: 21988367]
128. Srinivasan J, Kaplan F, Ajredini R, Zachariah C, Alborn HT, et al. 2008 A blend of small molecules regulates both mating and development in *Caenorhabditis elegans*. Nature 454:1115–18 [PubMed: 18650807]
129. Edison AS, Clendinen CS, Ajredini R, Beecher C, Ponce FV, Stupp GS. 2015 Metabolomics and natural-products strategies to study chemical ecology in nematodes. Integr. Comp. Biol 55:478–85 [PubMed: 26141866]
130. Choe A, Chuman T, von Reuss SH, Dossey AT, Yim JJ, et al. 2012 Sex-specific mating pheromones in the nematode Panagrellus redivivus. PNAS 109:20949–54 [PubMed: 23213209]
131. Wolfender J-L, Bohni N, Ndjoko-Ioset K, Edison AS. 2017 Advanced spectroscopic detectors for identification and quantification: nuclear magnetic resonance. In Liquid Chromatography: Fundamentals and Instrumentation, ed. Fanali S, Haddad PR, Poole CF, Schoenmakers F, Lloyd D, pp. 349–84. Amsterdam: Elsevier

132. Schroeder FC, Taggi AE, Gronquist M, Malik RU, Grant JB, et al. 2008 NMR-spectroscopic screening of spider venom reveals sulfated nucleosides as major components for the brown recluse and related species. PNAS 105:14283–87 [PubMed: 18794518]

133. Pungaliya C, Srinivasan J, Fox BW, Malik RU, Ludewig AH, et al. 2009 A shortcut to identifying small molecule signals that regulate behavior and development in *Caenorhabditis elegans*. PNAS 106:7708–13 [PubMed: 19346493]

134. Robinette SL, Brüschweiler R, Schroeder FC, Edison AS. 2012 NMR in metabolomics and natural products research: two sides of the same coin. Acc. Chem. Res 45:288–97 [PubMed: 21888316]

135. Bingol K, Brüschweiler-Li L, Yu C, Somogyi A, Zhang F, Brüschweiler R. 2015 Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. Anal. Chem 87:3864–70 [PubMed: 25674812]

136. Wang C, He L, Li DW, Brüschweiler-Li L, Marshall AG, Brüschweiler R. 2017 Accurate identification of unknown and known metabolic mixture components by combining 3D NMR with Fourier transform ion cyclotron resonance tandem mass spectrometry. J. Proteome Res 16:3774–86 [PubMed: 28795575]

137. Wang B, Dossey AT, Walse SS, Edison AS, Merz KM Jr. 2009 Relative configuration of natural products using NMR chemical shifts. J. Nat. Prod 72:709–13 [PubMed: 19265431]

138. 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:E31 [PubMed: 29748461]

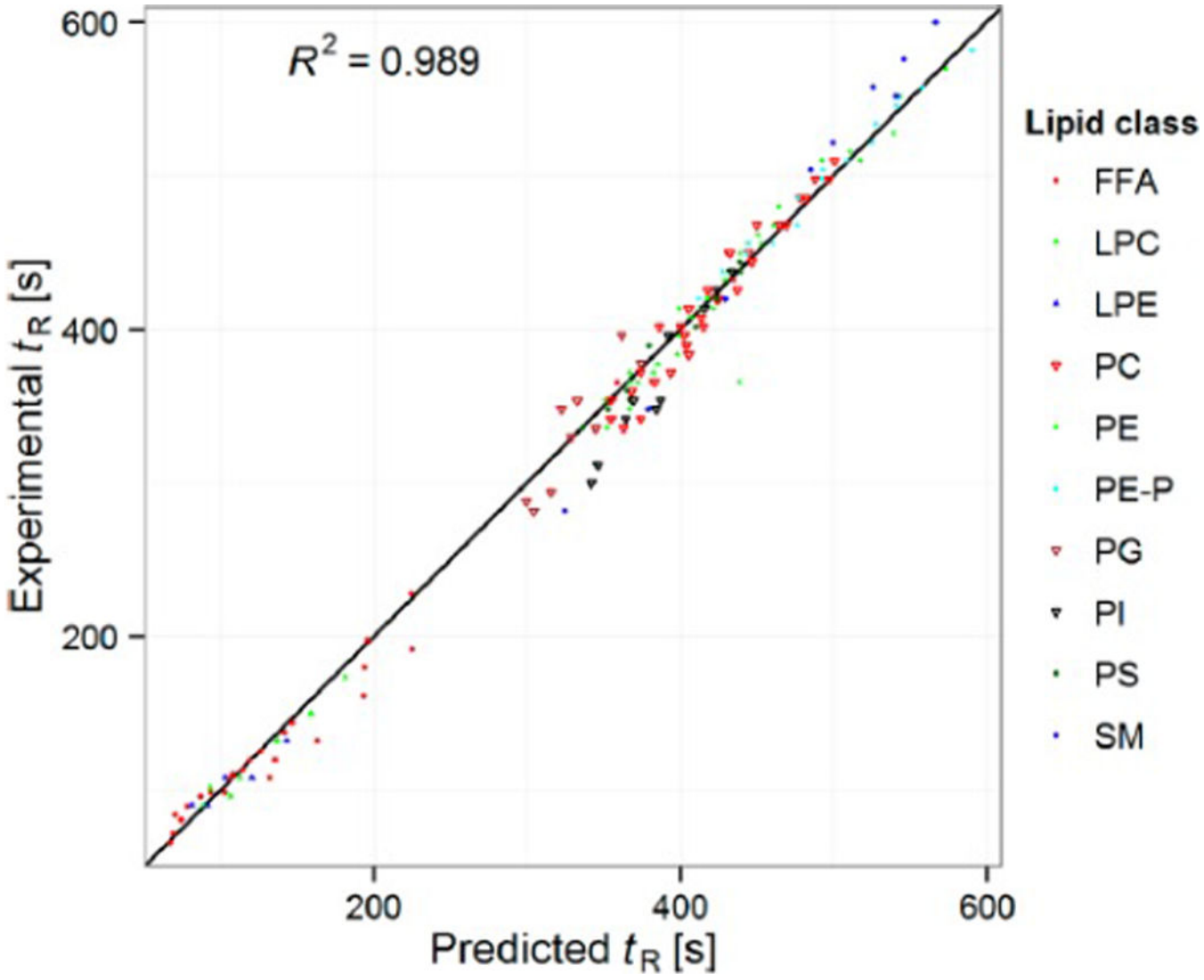

Figure 1.

Comparison of predicted and experimental retention times of the initial model. The model was trained on 50 data points and evaluated for 151 validation analytes. The depicted test lipids are distinguished by lipid class. The listed R^2 was computed from the test lipids. Reproduced from Ref. 33 with permission from the American Chemical Society ©2015.

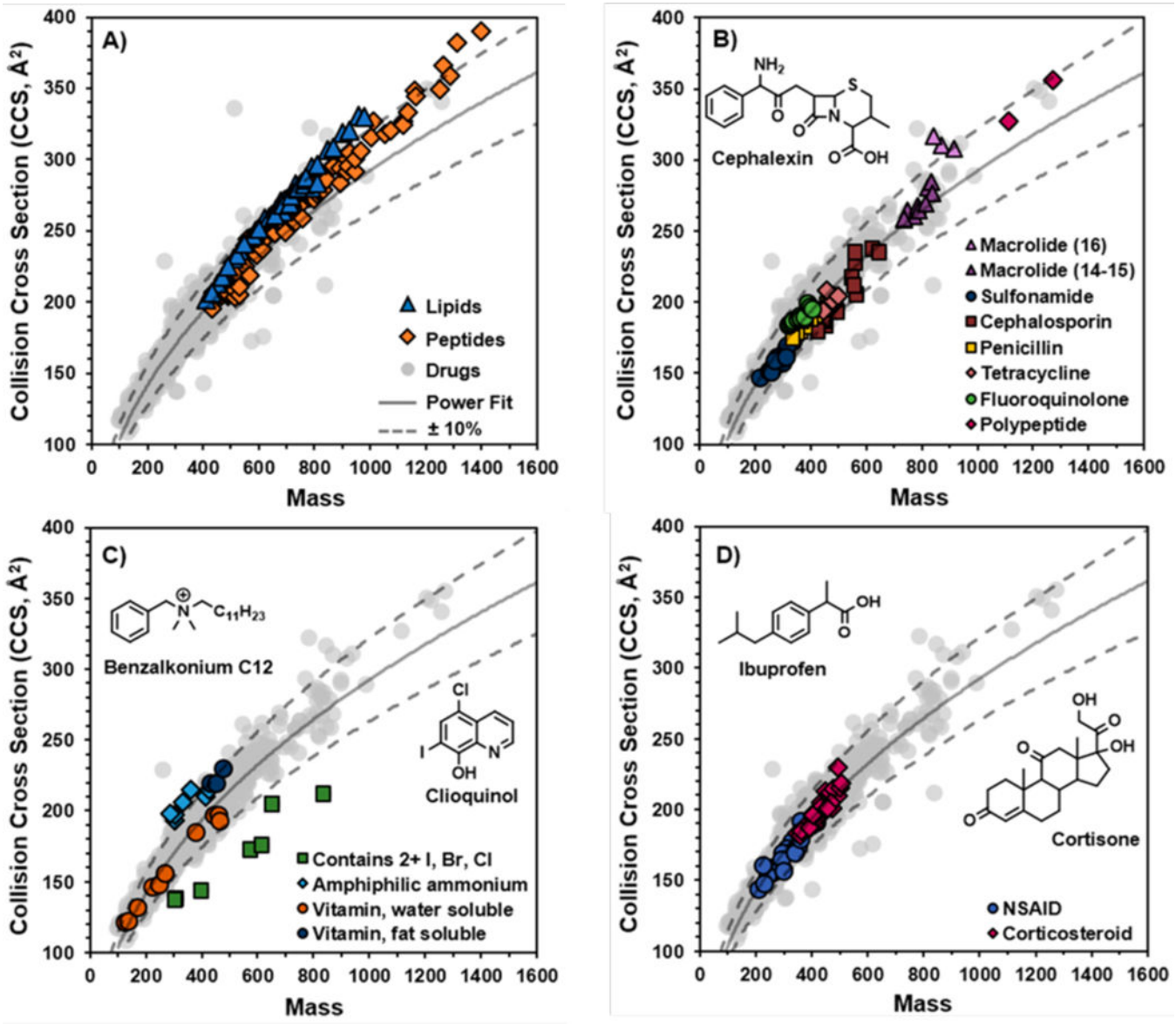

Figure 2.

IMS-MS plots showing the regions occupied by (A) lipids and peptides; (B) subclasses of antibiotics; (C) compounds of various densities; and (D) corticosteroid and nonsteroidal anti-inflammatory drugs (NSAIDS). Structures are shown for: cephalexin a cephalosporin antibiotic, benzalkonium C12 an amphiphilic ammonium, clioquinol an antifungal drug, ibuprofen a common NSAID, and cortisone a common corticosteroid. Reproduced from Ref. 100 with permission from the American Chemical Society ©2017.

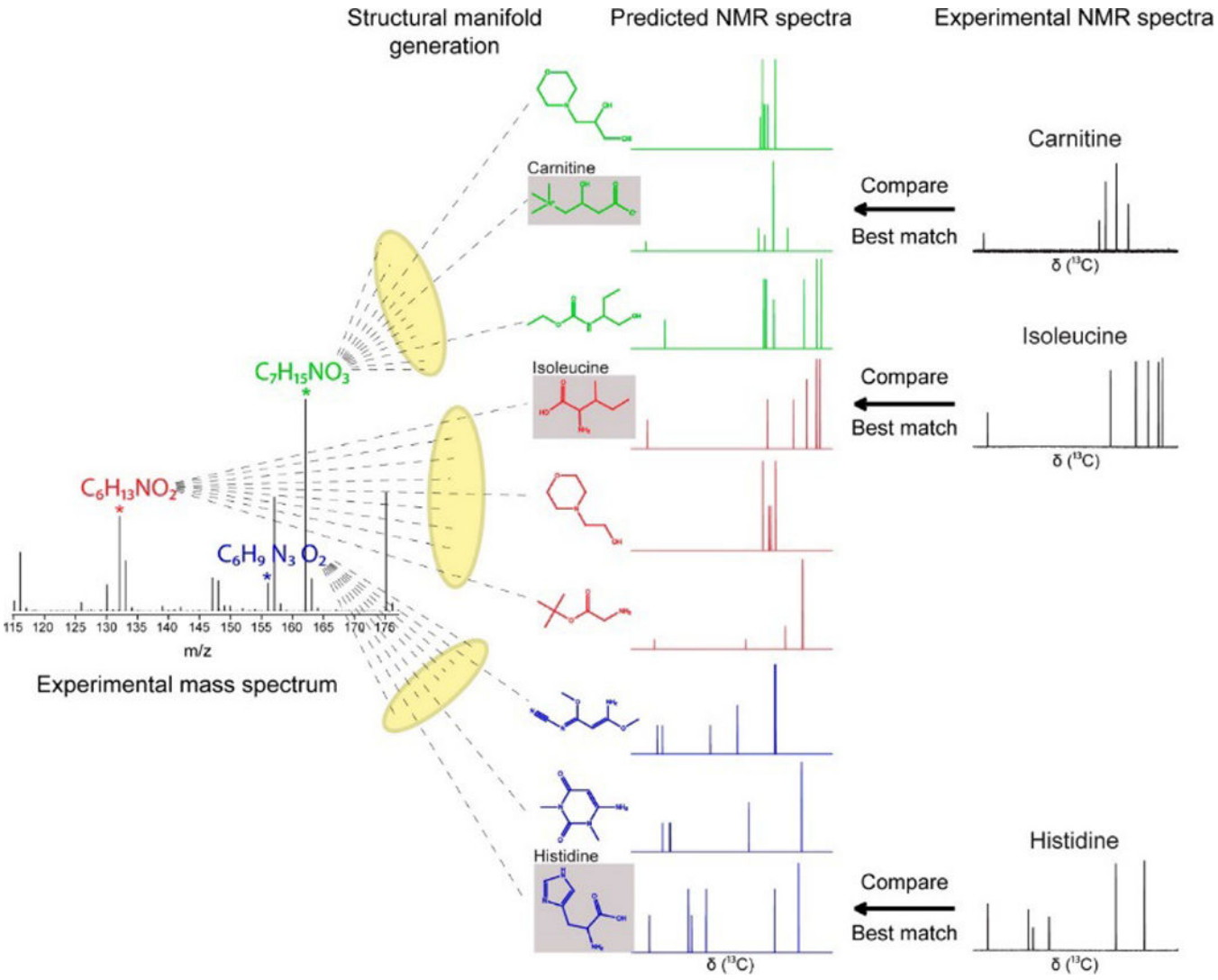

Figure 3.

The SUMMIT approach is a powerful way to link high-resolution MS and NMR data. Starting from an accurate mass measurement and molecular formula, NMR chemical shift calculations are done on all potential structures consistent with that formula. The computed NMR chemical shifts are then compared to experimental NMR data, and the best match is the most likely structure. Figure from Ref. 136 with permission from the American Chemical Society ©2015.

Table 1

Proposed “score card” approach to assign metabolite identity confidence levels. In this approach, points are added as various orthogonal methods of molecular characterization are successfully applied. A higher score indicates higher certainty on metabolite identity.

Qualifier		Explanation	Points assigned
(i) High resolution MS *	a. Monoisotopic peak *m/z* match within 5–10 ppm.	Simplest metric for assigning a tentative identity. Should be used with caution, as high ambiguity exists in ID assignment based solely on *m/z*.	5
	b. Monoisotopic peak *m/z* match within 1–5ppm.	Higher mass accuracy reduces the number of elemental formula candidates, helping in reducing the number of tentative IDs.	10
	c. Mass-to-charge ratio match within 1–2 ppm and isotopic ratio match with proposed elemental formula.	Relative abundance information found in isotopic structure can further reduce number of plausible elemental formula candidates.[1]	15
	d. More than one matched ESI adduct or in-source fragment (e.g. $[M-H]^-$, $[M+Cl]^-$).	The match of elemental formulae to more than one spectral feature (e.g. $[M+H]^+$ and $[2M+H]^+$) results in increased confidence.	5
	e. Mass-to-charge ratio match within 1–2 ppm and X+1, X+2 ion isotopic fine structure (e.g. high field Orbitrap, FT ICR).	Ultrahigh resolution MS measurements allow readout of elemental formulae directly from mass spectrum.	20
		Sub-Total Max	**25**
(ii) MS^N	Match of MS^2 fragmentation spectrum to experimental, database, or literature spectra.	Tandem MS experiments, although not 100% conclusive, greatly increase identification confidence.	10 total, proportional to # fragments matched
	Match of expected fragment ion ratios	Relative ratios of fragment ion pairs can discriminate between closely related, isobaric, species	5
	Match of experimental MS^2 fragmentation spectrum to *in silico* predictions	If database entries are not available, or not match is found, some degree of success can be obtained by predicting CID fragmentation via software tools (e.g. Thermo Mass Frontier, CFM-ID, MetFrag, Mass Fragment)	5
	Manual interpretation of MS^2 fragmentation spectrum consistent with proposed structure.	Manual interpretation of tandem MS spectra is sometimes necessary in the absence of database matches or unsuccessful *in silico* prediction.	10
	Higher than MS^2-level match to MS^N database (such as *mzcloud*)	Emerging databases now include MS^N information.	5
		Sub-Total Max	**20**
(iii) Chromatography	Retention time match (within expected window) between candidate ID and chemical standard.	Chromatographic retention times are excellent qualifiers for increasing ID certainty. Analytes should elute outside of the dead volume window.	10 (+5 if spiked in sample)
	In absence of standard, retention time may be matched to predicted value.	QSRR tools to predict retention time for small metabolites can be used to reinforce ID confidence.	10
		Sub-Total Max	**20**
(iv) Ion mobility	Database CCS matches to experimental CCS.	Ion mobility is an emerging technique in terms of metabolite ID. Databases are being created to facilitate identification using CCS.	10
	Predicted CCS matches experimental CCS	Molecular descriptor-based prediction of CCS is possible.	5
	IM filtering post precursor ion selection prior to MS^2	Better matches of experimental MS^2data to database data can be obtained by IM separation prior to CID, mitigating precursor ion co-selection.	5

Qualifier		Explanation	Points assigned
		Sub-Total Max	**15**
(v) NMR	1D NMR match (e.g. ^{1}H or ^{13}C)	Matches to COLMAR or any other query interfaces	15
	2D NMR match	Matches to COLMAR or any other query interfaces/ databases.	20
‘		**Sub-Total Max**	**20**

* Only ia, ib or ic should be applied.

(1) Pluskal, T.; Uehara, T.; Yanagida, M. *Anal. Chem.* **2012**, 84, 4396–4403.