Introducing 'identification probability' for automated and transferable assessment of metabolite identification confidence in metabolomics and related studies

Thomas O. Metz[1*], Christine H. Chang[1], Vasuk Gautam[2], Afia Anjum[2], Siyang Tian[2], Fei Wang[3,4], Sean M. Colby[1], Jamie R. Nunez[1], Madison R. Blumer[1], Arthur S. Edison[5], Oliver Fiehn[6], Dean P. Jones[7], Shuzhao Li[8], Edward T. Morgan[9], Gary J. Patti[10], Dylan H. Ross[1], Madelyn R. Shapiro[11], Antony J. Williams[12], David S. Wishart[2]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA USA

[2]Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

[3]Department of Computing Science, University of Alberta, Edmonton, AB, Canada

[4]Alberta Machine Intelligence Institute, Edmonton, AB, Canada

[5]Department of Biochemistry & Molecular Biology, Complex Carbohydrate Research Center and Institute of Bioinformatics, University of Georgia, Athens, GA, USA

[6]West Coast Metabolomics Center, University of California Davis, Davis, CA, USA

[7]Clinical Biomarkers Laboratory, Department of Medicine, Emory University, Atlanta, Georgia, USA

[8]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

[9]Department of Pharmacology and Chemical Biology, Emory University School of Medicine, Atlanta, Georgia, USA

[10]Center for Mass Spectrometry and Metabolic Tracing, Department of Chemistry, Department of Medicine, Washington University, Saint Louis, Missouri, USA

[11]Artificial Intelligence & Data Analytics Division, Pacific Northwest National Laboratory, Richland, WA USA

[12]U.S. Environmental Protection Agency, Office of Research & Development, Center for Computational Toxicology & Exposure (CCTE), Research Triangle Park, NC USA

*Corresponding author: thomas.metz@pnnl.gov

Author ORCID

Thomas O. Metz – 0000-0001-6049-3968

Christine H. Chang

Vasuk Gautam

Afia Anjum

Siyang Tian

Fei Wang – 0000-0002-0191-9719

Sean M. Colby

Jamie R. Nunez

Madison R. Blumer

Art S. Edison – 0000-0002-5686-2350

Oliver Fiehn – 0000-0002-6261-8928

Dean P. Jones – 0000-0002-2090-0677

Shuzhao Li – 0000-0002-7386-2539

Edward T. Morgan – 0000-0003-4273-2261

Gary J. Patti – 0000-0002-3748-6193

Dylan H. Ross – 0009-0005-2943-2282

Madelyn R. Shapiro

Antony J. Williams – 0000-0002-2668-4821

David S. Wishart – 0000-0002-3207-2434

ABSTRACT

Methods for assessing compound identification confidence in metabolomics and related studies have been debated and actively researched for the past two decades. The earliest effort in 2007 focused primarily on mass spectrometry and nuclear magnetic resonance spectroscopy and resulted in four recommended levels of metabolite identification confidence – the Metabolite Standards Initiative (MSI) Levels. In 2014, the original MSI Levels were expanded to five levels (including two sublevels) to facilitate communication of compound identification confidence in high resolution mass spectrometry studies. Further refinement in identification levels have occurred, for example to accommodate use of ion mobility spectrometry in metabolomics workflows, and alternate approaches to communicate compound identification confidence also have been developed based on identification points schema. However, neither qualitative levels of identification confidence nor quantitative scoring systems address the degree of ambiguity in compound identifications in context of the chemical space being considered, are easily automated, or are transferable between analytical platforms. In this perspective, we propose that the metabolomics and related communities consider identification probability as an approach for automated and transferable assessment of compound identification and ambiguity in metabolomics and related studies. Identification probability is defined simply as 1/N, where N is the number of compounds in a reference library or chemical space that match to an experimentally measured molecule within user-defined measurement precision(s), for example mass measurement or retention time accuracy, etc. We demonstrate the utility of identification probability in an *in silico* analysis of multi-property reference libraries constructed from the Human Metabolome Database and computational property predictions, provide guidance to the community in transparent implementation of the concept, and invite the community to further evaluate this concept in parallel with their current preferred methods for assessing metabolite identification confidence.

INTRODUCTION

Comparing Molecular Identification Among Omics Measurements

In biomedical research, systems biology studies[1-3] are used to discover new disease biomarkers and elucidate underlying biological mechanisms. Such studies are driven by multiple high-throughput omics technologies: genomics,[4, 5] transcriptomics,[6] proteomics[7, 8] and metabolomics.[9, 10] Genomics and transcriptomics are the most mature, owing to the more limited chemical diversity of DNA and RNA relative to proteins or metabolites,[11, 12] the fidelity and accuracy of the associated measurement techniques (i.e. sequencing)[5], and the breakthrough of having a complete human genome reference sequence as a result of the Human Genome Project.[13] Today, whole genomes can be sequenced in just 1-2 days with error rates <0.1%,[14] using modern high throughput sequencing technology (e.g. Illumina NovaSeq) and exploiting the fidelity of DNA polymerase for molecular replication and the specificity of fluorophores read from labeled base pairs.[5]

Proteomics is next in technical maturity. This is because proteins have only slightly greater chemical diversity compared to DNA and RNA, as they are composed of 22 amino acids. However, the complexity of the proteome can increase greatly if all possible protein post-translational modifications (PTMs; e.g. phosphorylation) are considered, and the computational time required for processing mass spectrometry-based proteomics data scales exponentially with the number of PTMs considered. Mass spectrometry-based proteomics[7, 8] exploits several characteristics of proteins and their constituent peptides. First, proteins are direct readouts of the genetic code, and if the genome is known, then associated protein sequences can be determined.[15] Second, peptides dissociate characteristically around the amide bond during a tandem mass spectrometry (MS/MS) measurement, allowing for accurate prediction of their fragmentation spectra.[16, 17] These characteristics have led to analytical workflows that can determine the proteomes of moderately complex samples, as well as methods for estimating and controlling peptide and protein identification false discovery rates (FDRs).[18, 19] Completely measuring the proteomes of highly complex samples (e.g., human blood plasma) requires a balancing of time and cost. In addition, comprehensive determination of post-translationally modified proteins[20] and hybrid peptides[21] remains challenging.

Metabolomics is the least mature among the omics sciences, with high-throughput, untargeted measurements having the goal of identifying and quantifying as many non-protein, small molecules (e.g., 50-1500 Da) as possible. Given their high sensitivity and broad molecular coverage, a variety of liquid chromatography-mass spectrometry (LC-MS) techniques are used in untargeted metabolomics. Typically, LC-MS assays yield thousands of signals with unique *m/z* and retention time (RT) coordinates. Each signal, defined as a “feature”, represents a potential small molecule of interest. However, these features may also be due to chemical noise or contamination and chemical variants of small molecules such as protonated- or sodiated-adducts. The rate-limiting step in untargeted metabolomics is discerning among these signals to annotate the chemical structures associated with the detected features. The current paradigm for confident metabolite identification involves comparing experimental MS (or nuclear magnetic resonance spectroscopy; NMR) data from biological measurements to comparable data from purified reference metabolites that were measured under similar conditions, preferably in the same laboratory. Unlike proteomics, where the analytes of interest are encoded by the genome and limited to linear polymers of repeating amino acids, the chemical space being profiled in metabolomics is essentially unconstrained, especially if exogenous metabolites (such as food products), microbial transformations and other chemical exposures are considered. As a result, much less is known about the complete composition of the human metabolome than the genome or the proteome. This is because relatively few reference standards are available relative to the known chemical space, and the measured properties such as mass fragmentation patterns are less predictable for metabolites than peptides. Consequently, metabolite identification in metabolomics is often prone to more errors or uncertainties than other omics technologies. Even if we could computationally predict all metabolites likely to exist in a given organism or biofluid, based on genomes or proteomes, the search would not be complete due to interaction of the organism with non-biological sources. As a result, even though metabolomics reference libraries continue to grow[22, 23], they are unlikely to ever be complete. This has inspired efforts to increase reference data through enzymatic biotransformation of drugs and other xenobiotic chemicals.[24]

Placing Confidence in Metabolomics Identifications

Insufficient knowledge of, or constraints placed upon, which small molecules might be present in a sample creates unique challenges when attempting to identify the chemical structure associated with a feature detected in metabolomics analyses. Even with the most recent developments in software and innovative computational methods that can automate steps in the informatics workflow,[25, 26] a critical question is the level of confidence that one has in the identifications proposed. The extent to which experimental data collected from a sample matches reference data is typically used to support feature identification. Unfortunately, there are often dozens to thousands of possible isomers in chemical and metabolomic databases. Isomers are chemicals with the same elemental formulae but different three-dimensional structures or different atomic positions. MS alone is not capable of disambiguating most of these isomers. In addition, other kinds of isomers may exist for which reference data do not yet exist.[27] Hence, apart from the accurate mass (monoisotopic *m/z*) data and MS/MS spectra obtained from a MS measurement, complementary data from additional analytical measurements (e.g., retention time, collision cross section (CCS), NMR spectra, different ionization modes or chemical derivatizations) improve identification confidence by limiting the number of potential compounds that satisfy the given match criteria[28]. However, currently there is no method for quantifying the ambiguity in a metabolite identification in context of the chemical space being considered. Accurately estimating total FDR in compound identifications is still in its infancy in metabolomics.[29-31]

In 2005, a Metabolomics Standards Workshop[32] was convened by the U.S. National Institutes of Health and the Metabolomics Society with the goal of establishing a Metabolomics Standards Initiative (MSI)[33] that would consider and recommend minimum reporting standards for describing various aspects of metabolomics experiments. The MSI consisted of five working groups comprised of international experts in metabolomics research and that developed recommended requirements for biological context, chemical analysis, data processing, ontology, and data exchange associated with metabolomics studies. In 2007, the Chemical Analysis Working Group of the MSI published the seminal paper on the minimum information for reporting the chemical analysis metadata associated with a metabolomics study, including a 4-level, qualitative scheme for reporting metabolite identification confidence.[28] These MSI-levels have been revised to include additional considerations[34] or other data types[35] but have remained largely unchanged. In 2014, Sumner *et al.*[36] and Creek *et al.*[37] proposed a transition from the existing qualitative metabolite identification confidence levels to a quantitative scoring system based on identification points (IP), citing the bias of the traditional MSI-levels towards identifications made in the context of data from authentic reference compounds or the need for more granularity in the levels, respectively. Most recently, Alygizakis and colleagues used a machine learning approach to develop a new IP-based system.[38] All three papers cited the EU Guideline 2002/657/EC[39] as a motivating example.

Reporting qualitative MSI-confidence levels in metabolite identifications is infrequently and inconsistently used by members of the metabolomics community. This is likely because assigning confidence scores is still a subjective process for most data reporters. Recipients of such data reports lack sufficient information or tools to independently verify metabolite identifications. Many reports include only chemical names, but not chemical or structure identifiers like PubChem Compound Identifications (PubChem CIDs) or International Chemical Identifiers (InChIs).[40] Chemical names can be highly ambiguous and misleading for data consumers and easily lead to problems in comparing data across different biological studies, as recently highlighted by arguments in the lipidomics literature.[41] For scientists who process LC-MS/MS data, deciding whether a given experimental MS/MS spectrum matches a reference spectrum is dependent on the metric used, the threshold set, and many other ambiguous decision points.[29] The current best alternative to MSI-confidence levels is to provide both raw and processed data in public repositories such as the Metabolomics Workbench[42] and MetaboLights[43] to support claims of reported metabolite identifications and to allow for independent verification.

Expanding from Measures of Confidence to Measures of Ambiguity

For both qualitative levels of metabolite identification confidence[28, 34, 35] and quantitative scoring systems[36, 37], the methods are not easily automated or transferable between analytical platforms (e.g., MS and NMR) and the degree of ambiguity or uncertainty in identifications is not fully represented. That is, given a reference library of a certain size and composition, and an analytical approach of certain resolution and precision, what is the likelihood of one identification being more correct than another given the available evidence? Here, we introduce a concept for moving from levels of identification confidence or cumulative point scoring systems to a universal method that assigns a mathematical probability to a given identification being correct. Importantly, this concept considers the composition and size of the reference library used, the numbers and types of measurement dimensions included in the experimental analysis, and each measurement's precision. It is also easily automated and the results transferable between analytical platforms.

INTRODUCTION TO METABOLITE IDENTIFICATION PROBABILITY

Logic Supporting the Concept

Metabolite identification probability represents a first step in moving away from semi-manually assigning subjective, constantly evolving levels of identification confidence (e.g., MSI levels) or IP-based methods towards a universal, automated method. Importantly, while methods for estimating FDRs for non-peptide small molecules have been explored in the context of MS/MS spectral matching, these have not been extended to other technologies (e.g., NMR) and data types (e.g., retention times, CCS values). The identification probability concept that we introduce here can be applied to any metabolomics measurement technology or method that relies on reference libraries (e.g., MS, GC-MS, LC-MS, LC-ion mobility spectrometry (IMS), LC-IMS-MS, LC-IMS-MS/MS, NMR, LC-NMR, etc.). Identification probability is defined as follows:

$$\text{Identification Probability} = 1/N$$

where N is the number of molecules in a reference library that match an experimentally measured feature within the precision(s) of the given measurement technology or method and the user-defined tolerances allowed in the measurement precision(s)

Based on this definition, higher dimensional analytical approaches or those that provide measurements of more properties should provide higher probability in a compound identification due to their ability to provide higher resolution of chemical space, while larger reference libraries would make it more difficult to completely resolve molecules in chemical space due to higher potential for conflicts.

Let's consider a single dimension or single property analysis to start. MS when used alone produces mass spectra, and the spectra will have a given resolution, based on the type of mass spectrometer used. Fourier transform ion cyclotron resonance (FTICR)-MS provides the highest mass resolution among current mass spectrometers used for metabolomics and related studies and can lead to extremely high accuracy in determining the exact molecular formulae that correspond to detected isotope patterns in the mass spectrum. The determined molecular formulae can then be searched against an appropriate reference library consisting of known molecular formulae; in our example, we consider a subset of the Human Metabolome Database (HMDB)[44] consisting of 22,077 non-lipid molecules for which computationally-predicted reference data were generated (**Supplemental Table S1**). If one were to perform a metabolomics experiment and detect a feature with protonated exact mass equal to 116.07115 Daltons, then the calculated molecular formula would be $C_5H_9NO_2$, which may correspond to the target molecule 4-amino-2-methylenebutanoic acid. When that formula is searched against the down-selected HMDB library that contains molecular formulae for up to 9 compounds with the same formula, then the probability of the experimentally measured formula $C_5H_9NO_2$ actually being 4-amino-2-methylenebutanoic acid (or any of the 9 candidates) would be 1/9 or 11% (**Figure 1**). Now, on the other end of the extreme, let's consider a multi-dimensional analysis, such as IMS-MS/MS. From this analysis, we would determine an IMS drift time or CCS value, a MS/MS spectrum and an accurate mass. The individual measurement precisions of any of these dimensions is not sufficiently high as to allow exact determination of any given property, and so matching of experimental data to the library proceeds within ranges or tolerances determined by typical experimental precision: ± 10 ppm for mass, ± 1% for CCS, and ≥ 850 for cosine similarity score (for MS/MS spectral matching). In the example shown in **Figure 1** for the target molecule 4-amino-2-methylenebutanoic acid, the combination of ± 10 ppm and ± 1% CCS reduces the candidates in the reference library to 7, and the identification probability for all candidates is 1/7 or 14%. For the same example, the combination of ± 10 ppm, ± 1% CCS, and ≥ 850 cosine similarity score reduces the candidates in the reference library to 1, and the identification probability is 1/1 or 100% for the measured feature corresponding to the target molecule 4-amino-2-methylenebutanoic acid. A key advantage to higher dimensional analyses is that the likelihood in complete overlap among property sets for library entries decreases roughly in proportion to the number of dimensions of the analysis.

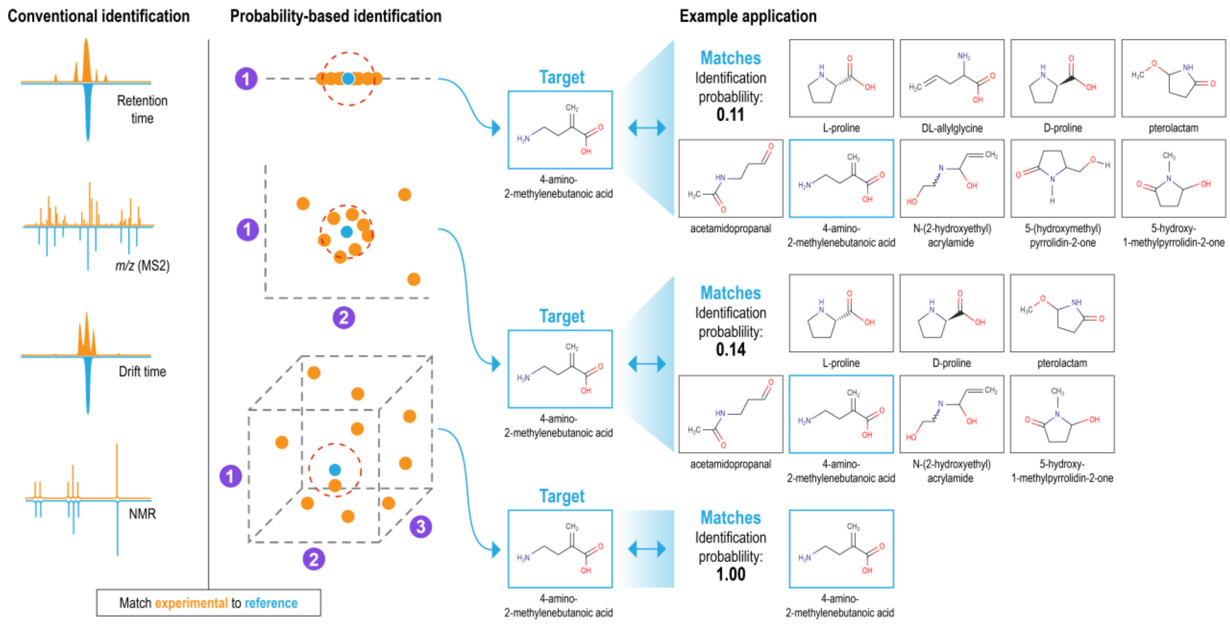

Conventional identification
Probability-based identification
Example application
Retention time
m/z (MS2)
Drift time
NMR
Match experimental to reference
1
2
3
Target
4-amino-2-methylenebutanoic acid
Matches
Identification probability: 0.11
L-proline
DL-allylglycine
D-proline
pterolactam
acetamidopropanal
4-amino-2-methylenebutanoic acid
N-(2-hydroxyethyl) acrylamide
5-(hydroxymethyl) pyrrolidin-2-one
5-hydroxy-1-methylpyrrolidin-2-one
Target
4-amino-2-methylenebutanoic acid
Matches
Identification probability: 0.14
L-proline
D-proline
pterolactam
acetamidopropanal
4-amino-2-methylenebutanoic acid
N-(2-hydroxyethyl) acrylamide
5-hydroxy-1-methylpyrrolidin-2-one
Target
4-amino-2-methylenebutanoic acid
Matches
Identification probability: 1.00
4-amino-2-methylenebutanoic acid

Figure 1. Demonstration of the metabolite identification probability concept using 4-amino-2-methylenebutanoic acid as the target molecule. Conventional metabolite identification (left panel) is based on manual or semi-automated comparison of experimental data to similar data contained in reference libraries, with final identification confidence determined by a data analyst. Probability-based identification (right panel) is similarly based on comparison of experimental and reference library data and is fully automated. Identification probability is defined as 1/N, where N is the number of molecules in the reference library that match an experimentally measured feature within the precision(s) of the given measurement technology or method and the user-defined tolerances allowed in the measurement precision(s). In the examples shown, the target molecule is 4-amino-2-methylenebutanoic acid, and the reference library is a subset of HMDB consisting of 22,007 non-lipid molecules. In the top row, identification is based on a single dimension of analysis, formula match (1). In the middle row, identification is based on the combination of ± 10 ppm and ± 1% CCS matching (2). In the bottom row, identification is based on the combination of ± 10 ppm, ± 1% CCS, and ≥ 850 cosine similarity score match (3).

Impacts of reference library size, property match tolerances, and analysis dimensionality

To evaluate how library size, property match tolerances, and dimensionality of analytical analysis might impact metabolite identification probabilities, we further explored the 22,077 non-lipid molecules from HMDB, as well as a complementary set of 44,537 lipid molecules (**Supplemental Table S2**), from the same source. The molecules were classified into a chemical ontology using the ClassyFire tool,[45] and compounds with an invalid chemical classification value ("NA") were excluded (**Supplemental Figure S1**). The two molecule sets were placed in separate matrices, together with the protonated mass (*m/z*), RT, CCS, and MS/MS spectra for each molecule. The protonated mass was calculated from the protonated molecular formula. DarkChem[46] was used to predict CCS for all lipid molecules; for non-lipid molecules, DarkChem, AllCCS,[47] and DeepCCS[48] were used to predict CCS for 9308, 10,669, and 2100 molecules, respectively, as indicated in **Supplemental Table S1**. RTs were predicted using Retip[49] under hydrophilic interaction liquid chromatography (HILIC) conditions for non-lipid molecules or reversed-phase chromatography conditions for lipid molecules (based on ClassyFire assigned superclass of "Lipids or Lipid-like molecules"), and MS/MS spectra were predicted using CFM-ID 4.0[50] at a "medium" collision energy level of 20 eV. We then matched each of the two molecule sets and their calculated/predicted properties to themselves to simulate the processing of a metabolomics data set.

Impact of reference library size. To evaluate the impact of reference library size on metabolite identification probability, we performed Monte Carlo simulations to randomly draw smaller library subsets (e.g. 1,000, 5,000, or 10,000 molecules) from the full lipids and non-lipids libraries. We evaluated 100 randomly drawn subsets for each library size, matched each subset to itself by mass (±10 ppm), aggregated results, and compared the number of matches returned per database search (Error! Reference source not found.**2**). Our results demonstrate that as the size o f the library increases, the relative proportion of matches at a given probability decreases; thus, smaller reference libraries will tend to yield artificially high identification probabilities. Comparing lipids vs non-lipids, the impact of reference library size on identification probability is more pronounced for libraries with more heterogenous content.

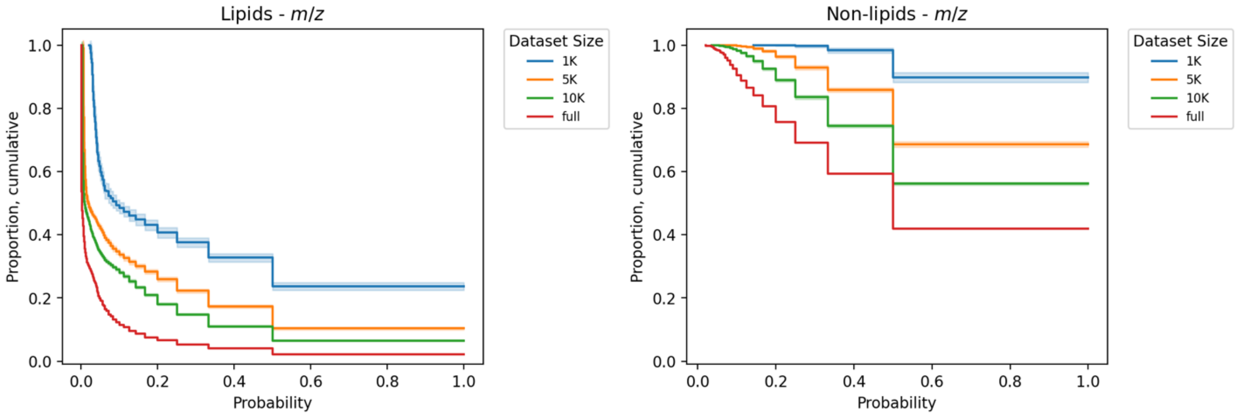
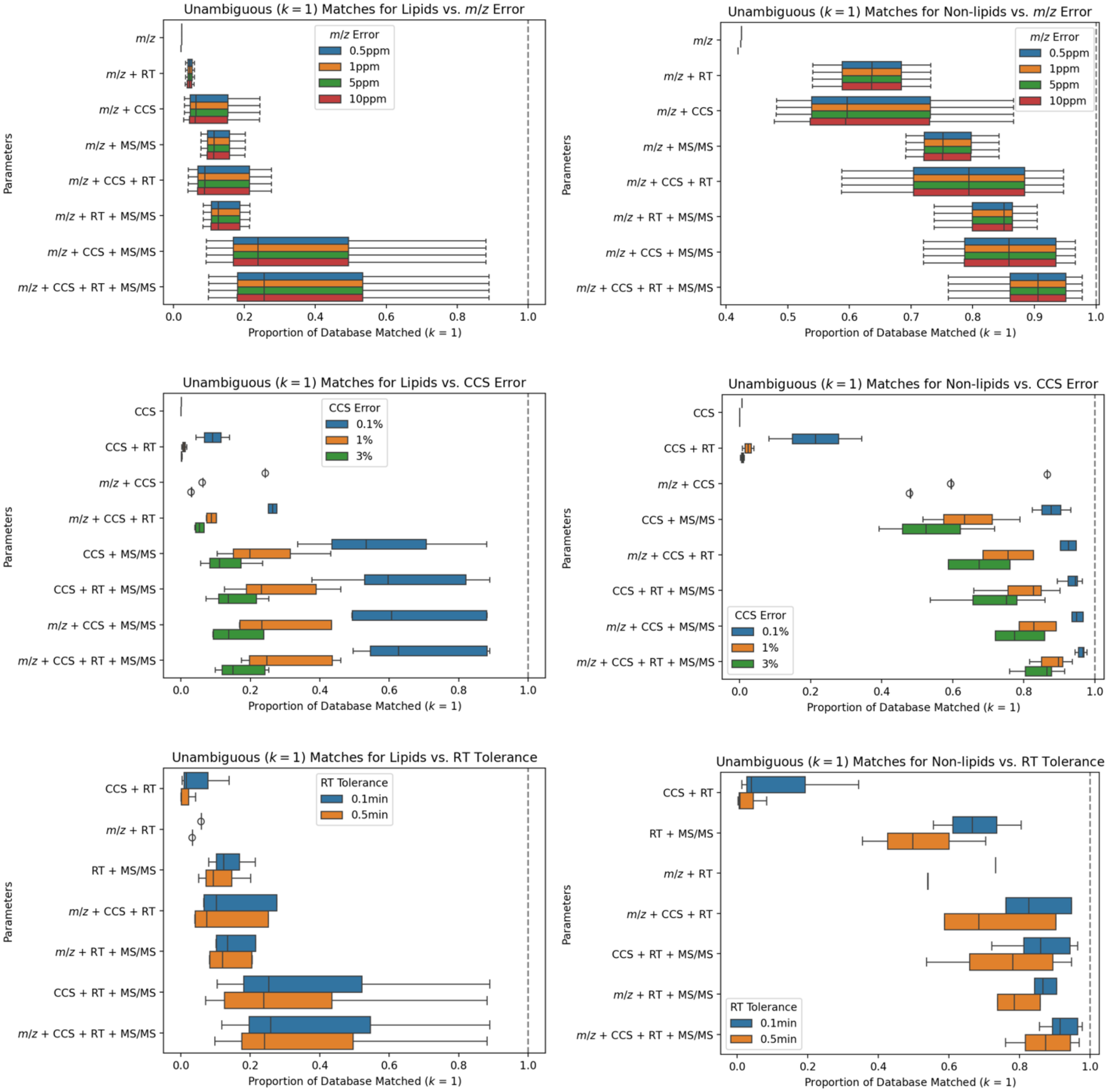


Figure 2. Impact of reference library size on metabolite identification probability. Monte Carlo simulations were performed to randomly draw subsets of the full lipids (left panel) and non-lipids (right panel) reference libraries of size 1K, 5K, and 10K. Match probability is shown on the x-axis, and the proportion of compounds in each dataset matched within ±10 ppm and with a given probability is shown on the y-axis. For example, for non-lipids, a little over 40% of compounds are unambiguously matched when matching the full library to itself with a mass tolerance of ± 10 ppm. Solid lines indicate the mean value, and shaded regions indicate ± 1 standard deviation from the mean based on 100 Monte Carlo simulations (note that no shaded region exists for the full dataset, for which random subsets were not drawn).

Impact of property match tolerances and dimensionality of analytical analysis. We next evaluated the impacts of individual property match tolerances and the dimensionality of the analytical analysis on metabolite identification probability. Overall, varying property match tolerance has different impacts on the number of unambiguous identifications depending on the property considered. For instance, the evaluated *m/z* match thresholds gave rise to little, if any, change in the proportion of unambiguous matches from both the lipids and non-lipids datasets, either alone or in combination with other properties (Error! Reference source not found.**3**). We h ypothesize that the low variance in match performance across *m/z* tolerances can be attributed to the relative density of compounds occupying *m/z* space vs. the variability of the error thresholds in practical terms. For instance, at an *m/z* of 800 Da (close to the median *m/z* for lipids of 821.8 Da), the error thresholds of ± 0.1, 1, 5, and 10 ppm correspond to ± 0.00008 Da, ± 0.0008 Da, ± 0.004 Da, and ± 0.008 Da, respectively. The resolutions may not differ sufficiently to effect significant changes to the number of matches within each corresponding tolerance.

In contrast to *m/z*, CCS search tolerance has a more pronounced impact on unambiguous matches. While searching by CCS alone produces zero or near-zero unambiguous matches across both lipids and non-lipids datasets, when used in combination with other analytical dimensions, the effect of CCS search tolerance becomes much more pronounced. In some cases,



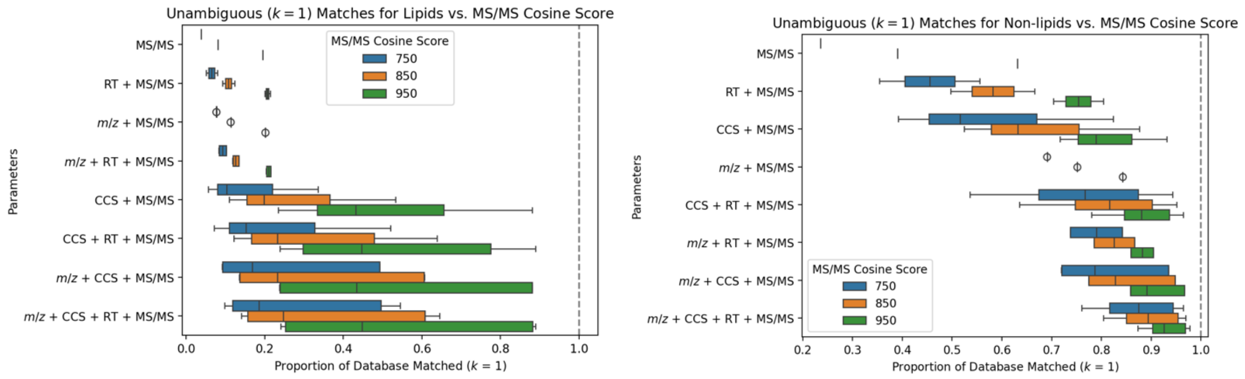


Figure 3. Impact of property match tolerances and dimensionality of analytical analysis on metabolite identification probability for lipids (left) and non-lipids (right). Each boxplot summarizes the fraction of each database that is unambiguously matched (*k* = 1) when varying the search tolerances evaluated in each dimension (*m/z*, CCS, RT, and MS/MS, respectively) as shown. The first set of boxplots in each plot represent results when only considering the dimension of interest and varying search tolerance within that single dimension, with the subsequent boxplots depicting results upon inclusion of additional search dimensions but only varying the search tolerance of the first dimension. For each dimension, search tolerances include *m/z* ± 0.1 ppm, ± 1 ppm, ± 5 ppm, and ± 10 ppm; CCS ± 0.1%, ± 1%, and ± 3%; RT ± 0.1 min, and ± 0.5 min; and MS/MS cosine score ≥ 750, ≥ 850, and ≥ 950.

we observe a two-fold or even greater increase in the fraction of the dataset which can be definitively matched, particularly in the case of lipids (Error! Reference source not found.**3**). Our s imulation data suggests that when used in conjunction with other measurements, accurate CCS measurements have the potential to increase the number of confident identifications. However, we note that the highest-accuracy CCS error threshold evaluated is a CCS error of ± 0.1%, which may be achievable experimentally only using very high-resolution ion mobility separations, such as structures for lossless ion manipulations (SLIM).[51] Inclusion of CCS in compound matching at this tighter threshold produced marked improvements in the proportion of unambiguous matches.

Neither of the two RT thresholds evaluated (± 0.1 min and ± 0.5 min) produced any unambiguous matches in the database using RT alone for the lipids or non-lipids datasets (data not shown). However, as with CCS, RT combined with additional measurement dimensions produced more confident matches (**Figure 3**). Reducing the RT tolerance from 0.5 min to 0.1 min correspondingly increases the proportion of unambiguous matches. While the observed effect is smaller than the impact of CCS, the inclusion of RT still substantially improves unambiguous identifications, especially compared with *m/z*.

Finally, we evaluated the impact of the MS/MS spectral match threshold. We chose to use cosine similarity score due to its ubiquitous use; however, we note that alternative scoring algorithms, such as spectral entropy,[29] have demonstrated improvements over cosine similarity. Based on the range of typical scoring thresholds used for MS/MS matching, we evaluated cosine similarity thresholds of 750, 850, and 950. Our results show that among both lipids and non-lipids, MS/MS score alone is the best-performing singular measurement in terms of identifying compounds unambiguously (**Figure 3**). In contrast to *m/z*, however, increasing the MS/MS cosine score threshold resulted in significant increases to the proportion of compounds producing unambiguous matches in both the lipids and non-lipids libraries. In fact, when matching by MS/MS cosine score alone, 63% of non-lipids can be accurately matched with a cosine similarity score of ≥950, compared to just 24% with a cosine score of 750. As before, the MS/MS dimension can be combined with other measurement dimensions to achieve an even greater fraction of unambiguous identifications; in fact, all the best-performing multi-dimensional search parameter sets include MS/MS.

While the example data and toy metabolite identification probability analyses discussed above are LC-MS-centric, the concept is applicable for any workflow that produces metabolite identifications through matching experimental data to similar data in reference libraries, such as NMR and GC-MS. Indeed, many NMR spectral matching algorithms, such as those used in MagMet[52], Bayesil[53] and Chenomx[54], use concepts similar to the cosine similarity score used in MS/MS. Likewise, GC-MS uses equivalent concepts as LC-MS/MS for spectral matching.

THE ROLE OF REFERENCE LIBRARIES AND HOW TO POPULATE THEM

Current Landscape and Use of Reference Libraries for Compound Identification

The metabolite identification probability concept introduced here depends on the size and contents of the reference library used. As such, it is essential that reference libraries are populated and used correctly. In the following discussion, we describe the current landscape and use of reference libraries for compound identification and provide recommendations to the community for their use as it relates to metabolite identification probability. For the purposes of this discussion, we assume that the contents of these reference libraries are correct and accurate.

Reference libraries contain varying levels of curated information about compounds (e.g., structure, properties, and classifications). At a minimum, useful reference libraries contain compound structures in machine readable formats or public identifiers that map to chemical structures, alongside derived properties such as elemental formulae and exact monoisotopic masses. In particular, many reference libraries developed for use with specific analytical approaches contain measurable observables, such as observed precursor ions and MS/MS or NMR spectra. They also include experimental metadata that define these spectra, such as the type of instrument used or e.g., details of the MS/MS fragmentation method that was applied. For the case of high-resolution MS (HRMS), the data can be used to directly search against exact masses of known, expected, and even predicted chemical structures. If HRMS data accuracy of <0.002 Da is achieved, chemical formulae can be inferred using a variety of different software tools, especially if MS/MS and isotope ratio information is included.[55-59] Note that a mass resolving power of R<250,000 means that alternative formulae might still need to be considered.[60]

Many open-access reference libraries exist in the form of compound collections that contain mass, formula and structure information for millions of known, suspected or predicted compounds (**Table 1**). These include PubChem[61] which has nearly 110 million compounds, ChEMBL[62] with 2.1 million compounds, and the US-EPA CompTox Chemicals Dashboard[63] with 1.2 million compounds. All of these support mass and formula searching. However, they also include a large fraction (>99%) of anthropogenic molecules, making these libraries somewhat more suited for exposomics[64] or environmental testing studies and less suitable for traditional metabolomics studies that focus on physiological metabolites. A number of reference libraries exist that focus on storing only known biologically-related compounds. For example, the Human Metabolome Database (HMDB) now accounts for 248,097 compounds,[44] Lipid Maps[65] lists 45,684 compounds, KEGG[66] denotes 18,784 compounds, and MetaCyc[67] includes 16,861 compounds. These databases continue to expand in coverage and content and such databases are much more suitable for traditional metabolomics studies.

Library	Number of Compounds	URL	Citation
ChemSpider	>129,000,000	http://www.chemspider.com/	68
PubChem	>119,000,000	https://pubchem.ncbi.nlm.nih.gov/	69
CompTox Chemicals Dashboard	>1,200,000	https://comptox.epa.gov/dashboard/	63
RaMP-DB 2.0	>256,000	https://rampdb.nih.gov/	70
Human Metabolome Database (HMDB)	>250,000	https://hmdb.ca/	44
Metabolomics Workbench	>164,000	https://www.metabolomicsworkbench.org/	42
Chemical Entities of Biological Interest (ChEBI)	>160,000	https://www.ebi.ac.uk/chebi/	71
LipidMaps	>47,000	https://www.lipidmaps.org/databases/lmsd/overview	65
Natural Products Atlas	>33,000	https://www.npatlas.org/	72
Metabolights	>27,000	https://www.ebi.ac.uk/metabolights/index	43
Kyoto Encyclopedia of Genes and Genomes (KEGG)	>19,000	https://www.genome.jp/kegg/	66
MetaCyc	>18,000	https://metacyc.org/	67

Table 1. Compound collection reference libraries. These reference libraries function primarily as collections of compounds and include chemical structures, molecular formulae, masses and physicochemical properties, among other data.

While *m/z* or formula searching is relatively easy to perform, and the sizes of the reference libraries mentioned above are often very large, the reliability of these single parameter matches is often quite poor. Indeed, it is often possible to get hundreds of potential matches with a single *m/z,* or even formula, query (**Figure 4**).[73, 74] Additional "observable" information is needed to add specificity and increase confidence in tentative compound identifications.[28, 34] Generally, the most accessible and reproducible experimental measurements, beyond molecular weight, are spectral

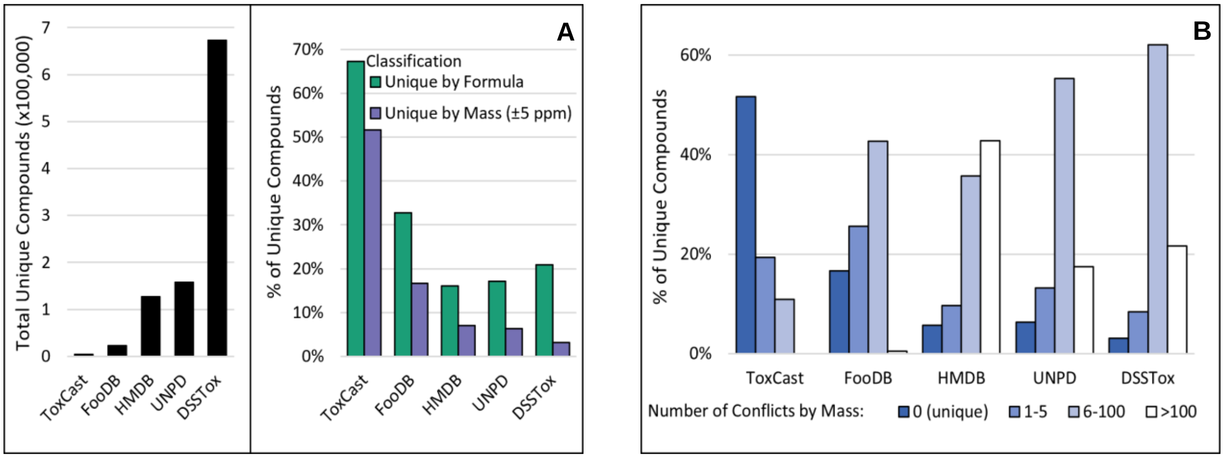


Figure 4. Size, composition, and uniqueness of representative reference libraries. (A) Number of unique compounds by structure (based on unique canonical SMILES generated by RDKit; Left Panel), and percent of compounds unique by formula or parent mass (Right Panel). Unique by parent mass indicates that there are no other compounds with a mass within 5 ppm. (B) Percent of library compounds that conflict by parent mass with up to >100 other molecules. Number of conflicts by parent mass is a count for how many other structures within the given library have a mass within 5 ppm. If there are 0 conflicts for a given structure, it is considered unique by parent mass.

or separations data. This includes MS/MS spectra (for LC-MS or CE-MS), electron ionization (EI) spectra (for GC-MS) or NMR spectra, and retention times (RT; for LC) or retention indices (RI; for GC) and drift times or collision cross sections (CCS) for IMS data. More recent technology developments allow for the collection of infrared spectra in-line with IMS and MS measurements.[75] The intensity, position, number and character of the peaks seen in MS/MS or NMR spectra is often considered sufficient to make identifications of metabolites; however, as shown in **Figure 3**, MS/MS spectral match alone is insufficient for providing unambiguous identification of metabolites when matching to large reference libraries. Several different scoring schemes are available to facilitate spectral matching and scoring and offer superior results to simply matching based on a mass or formula.[76, 77] Recently, spectral entropy was developed as a new MS/MS scoring scheme to particularly account for spectra with few fragment ions, as often observed in small molecule analyses.[29] The chromatographic and separation parameters are related to physicochemical properties (e.g., size, shape, charge, boiling point, hydrophobicity) and provide information that is fundamentally different from measured mass or fragmentation spectra. RI and CCS values can be relatively instrument- or condition-independent with proper calibration, making them highly reproducible and suitable for compound identification. CCS values are particularly reproducible, with relative standard deviations <1% reported in interlaboratory comparisons and under standardized conditions.[78] Fragmentation spectra (from GC-MS or LC-MS/MS) are generally relied upon the most in identification workflows due to their specificity and wide availability of associated instrumentation. GC-electron ionization mass spectra were standardized over 60 years ago. Yet, in comparison, measured spectra from LC-MS/MS are harder to standardize due to the variability between instruments, the fragmentation conditions and the collision energies used. Therefore, MS/MS libraries often contain multiple spectra for each compound.

Because of their utility in providing additional confidence in metabolite identification, there are a growing number of both commercial and open-access reference libraries that contain various properties from experimental measurements of pure reference compounds and that are available for matching to metabolomics data. Popular reference libraries that contain mass spectral data are MassBank.eu, MassBank of North America (MassBank.us), the NIST spectral library,[77] METLIN,[22] and mzCloud, as well as commercial libraries produced by Waters, Sciex, Bruker, Agilent, and Thermo Fisher. Other resources exist that contain both spectra from analysis of pure compounds but also large numbers of spectra of unknown compounds from analysis of real samples, such as GNPS.[79] Some of the more popular NMR spectral libraries are the BioMagResBank,[80] NMRShiftDB,[81] NP-MRD,[82] and COLMAR,[83] as well as commercial libraries produced by Bruker and Chenomx. Popular reference libraries that contain RI and/or CCS include: the NIST RI library, the FiehnLib RI library,[84] the Unified CCS Compendium,[85] the Sumner CCS library[86] and several commercial CCS libraries from instrument vendors such as Bruker, Agilent and Waters. MassBank.us contains many metabolites with LC-based retention times, including for hydrophilic interaction chromatography (HILIC)[49]. In contrast to standardized gas chromatography RI and CCS measurements, LC RT and electrophoretic mobilities are not easily translated from instrument to instrument or from one configuration to another. As a result, reference libraries for LC RT and electrophoretic mobility are often quite small. Recently however, the developers of METLIN released a reference library containing >80,000 RTs measured for small molecules, called SMRT.[87] These data, the largest of their kind, were collected using a single standard chromatographic protocol but has not been validated yet by independent means. A more detailed listing of reference libraries focused on housing data from analyses of pure reference compounds, their contents, and the number of entries found is provided in **Table 2**.

Library	**Number of Compounds**	**Number of Experimental Reference Values**	**URL**	**Citation**
Metlin	>860,000	5 million spectra	https://metlin.scripps.edu/	88
NIST20 EI-MS library	>306,000	>350,000 spectra	https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries	N/A
NIST RI library 2020	>139,000	>447,000 retention indices	https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:ridatabase	N/A
NIST20 Tandem MS library	>31,000	>1.3 million spectra	https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries	N/A
MassBank of North America (MoNA)	>227,000	>197,000 spectra	https://mona.fiehnlab.ucdavis.edu/	N/A
mzCloud	>21,000	>10.7 million spectra	https://www.mzcloud.org/	N/A
MassBank Europe	>15,000	>90,000 spectra	https://massbank.eu/MassBank/	N/A
Biological Magnetic Resonance Data Bank (BMRB)	>1,300	>10 million chemical shifts	https://bmrb.io/	80
NMRShiftDB	>40,000	>68,000 spectra	https://nmrshiftdb.nmr.uni-koeln.de/	81
Natural Product Magnetic Resonance Database (NP-MRD)	>87,000	>1500 spectra	https://np-mrd.org/	82
FiehnLib RI library	>1200	>1200 retention indices	https://fiehnlab.ucdavis.edu/projects/fiehnlib	84
AllCCS	>2100	>3500 CCS	http://allccs.zhulab.cn/	47
Unified Collision Cross Section Compendium	>1700	>3700 CCS	https://mcleanresearchgroup.shinyapps.io/CCS-Compendium/	85

Table 2. Reference libraries of observable data. These reference libraries contain listings of compounds and their observable data, such as mass spectra, retention indices, NMR spectra and CCS values.

Recent Advances in *In Silico* Tools for Expanding Reference Libraries

As can be seen from **Table 2**, measured observable data is very limited compared to the number of structures we know or suspect to exist. While many reference libraries containing experimentally determined values exist, most are currently too small or too incomplete to satisfy the needs of metabolomics studies. The most comprehensive untargeted MS-based metabolomics experiments that rely on today's reference libraries can identify up to 10% of the observed features.[89] However, such ratios depend on the type of data processing and assay: for GC-MS based metabolomics or in lipidomics assays, the ratio of identification is typically at 30% of features that have associated mass spectra.[90] In fact, high quality data processing should include measures of blank sample corrections, adduct deconvolution and the use of pooled sample quality controls to reduce the number of spurious features in assessments of metabolome coverage statements.

One route for increasing the amount of observable data in reference libraries is through the synthesis or isolation of molecules of interest. However, if one assumes that the total number of all known and predicted metabolites, as well as all known anthropogenic chemicals, found in humans is ~2 million compounds and the cost to isolate or synthesize and to comprehensively characterize these compounds is ~$5000/chemical, such an effort would cost in excess of $10 billion USD. This initiative would easily take 20+ years and consume a significant portion of the NSF or NIH budget. In other words, the time and cost to make the comprehensive reference library required for the metabolomics community is simply not feasible. A more cost-effective approach will have to be developed. We believe that a viable option, and the future of reference library growth, is via *in silico* approaches. Simply stated, computational approaches could be used to generate *in silico* (i.e., predicted) observable data, based on validated methods. We propose this because of the foundational developments in chemistry and physics and the need to identify a vast number of unidentified features. Development of various machine learning- or quantum chemistry-based approaches (reviewed in[91]) and tools for *in silico* prediction of various types of spectra and other observables has increased the size and chemical appropriateness of existing reference libraries. Indeed, there are now several well-developed software tools for predicting electron ionization-mass spectrometry (EI-MS), electrospray ionization-tandem mass spectrometry (ESI-MS/MS) and NMR spectra, CCS, and RT values using combinatorial approaches, machine and deep learning methods, and quantum mechanical techniques. For ESI-MS/MS spectral prediction, several machine learning methods including MetFrag,[92] CFM-ID,[93] MS-FINDER,[94] ChemDistiller[95] and MAGMa[96] have appeared. CFM-ID, MS-FINDER and MAGMa in particular have shown excellent performance in terms of spectral prediction accuracy in multiple independent tests.[97, 98] For EI-MS spectra, two machine learning methods (CFM-ID-EI[99] and NEIMS[100]) have been described and both perform well. Separately, a quantum mechanical method called QCEIMS[101] has been developed to predict EI-MS spectra and more recently ESI-MS/MS spectra with QCxMS.[102] QCEIMS and QCxMS are significantly slower than the ML methods, but they provide useful insights into the EI and ESI fragmentation processes.

Just as with EI-MS, both machine learning and quantum chemistry methods have been developed to predict NMR spectra (^{1}H, ^{13}C, 1D and 2D). Density Functional Theory (DFT) has been used for many years to predict NMR chemical shifts and coupling constants with errors as small as 0.2 ppm for ^{1}H shifts and 2.5 ppm for ^{13}C shifts.[103, 104] This level of precision can enable distinction of closely related diastereomers.[105] ISiCLE, which uses NWChem[106] for calculations, is an example of a recently developed DFT method that is now being used to calculate ^{1}H and ^{13}C NMR spectra for thousands of natural products that do not have measured NMR spectra in NP-MRD.[82] It is expected that ISiCLE will be able to create one of the world's largest *in silico*-predicted NMR spectral libraries by the end of 2024. It is also possible to use machine learning and neural networks to predict the NMR spectra of small molecules.[103, 107] These programs tend to be much faster than QM methods and may be just as accurate.[108]

Lastly, the prediction of separation properties, such as RI, RT, and CCS values, has become increasingly popular. Several machine learning-based programs for CCS prediction have recently appeared including DeepCCS,[48] MetCCS predictor,[109] and DarkChem.[46] Quantum chemical methods, such as ISiCLE,[110] have also been developed to accurately predict CCS values. Regardless of the method chosen, the typical errors between experimentally observed and predicted CCS values are as small as 2-3% with correlation coefficients greater than 0.95. Using these predictive CCS tools, several reference libraries have been generated, containing hundreds of thousands of predicted CCS values, including CCSBase,[111] AllCCS,[47] and MetCCS.[112] Similar efforts are being made in RI prediction and measured data curation. The NIST 20 library contains more than 114,000 experimentally measured Kovats retention indices, used for GC-based metabolomics. Using these data, NIST scientists have recently developed a graph neural network approach that can predict retention indices with a mean absolute percentage error (MAPE) as small as 3% and a correlation coefficient of >0.98.[113] This, by far, is the most accurate method for RI prediction ever published. A similarly accurate method for RI prediction has recently been implemented in the latest version of the HMDB which provides 6.7 million RIs for >26,000 GC-MS compatible compounds (and their derivatives).[44] In principle, these methods could be used to generate accurate RI values for each specific GC-MS method, for hundreds of thousands of molecules which do not have experimental RI data. In terms of RT prediction for LC, several efforts aimed at relative RT prediction have been undertaken using highly specified chromatographic conditions. These include the machine learning-based tools Retip,[49] GNN-RT,[114] and the METLIN SMRT predictor[87] that showed RT median prediction errors as small as 5%. However, the correlation coefficients between experimental and predicted RTs are often only ~0.6, suggesting that RT prediction for LC has a long way to go before it matches the accuracy of RI prediction.

With increasing popularity of *in silico* approaches to generating reference observables, there have been similar efforts to predict novel metabolite structures that can be added to reference libraries. Currently, there are two approaches for doing so. One is to use enzymatic modeling to predict biotransformations or enzymatic by-products of starting molecules.[115] The other is to use generative modeling and deep learning to create biofeasible structures.[46, 116] Biotransformation prediction has been around for many decades and was pioneered by researchers in the drug metabolism community.[117] As a result, a number of commercial programs have been developed, including Meteor Nexus, ADMET-Predictor, MetabolExpert and others, that predict Phase I (cytochrome P450) metabolism specifically for drug molecules and a small number of naturally occurring metabolites. These programs use expert-derived rules and large internal databases to perform look-ups and make their predictions. More recently, several open source or open access tools have appeared that perform biotransformation prediction for a larger collection of molecules. These include GLORYx,[118] FAME 2,[119] FAME 3,[120] CyProduct[121] and BioTransformer.[115] These software packages, many of which use machine learning techniques, not only predict Phase I biotransformation, but also Phase II metabolism and microbial/gut metabolism for drugs, pesticides, herbicides and naturally occurring metabolites. Furthermore, they are also able to predict these transformations much more accurately than commercial, rules-based software.

One of these programs, BioTransformer, has recently been applied to predict the structures of 2 million biotransformed molecules (Phase I + Phase II + microbial + promiscuous enzyme transformations) using a starting set of 120,000 compounds in the HMDB. Other approaches have also made use of enzyme promiscuity to predict biofeasible metabolites. For example, the MINE (Metabolic *In silico* Network Expansions) database used an algorithm called the Biochemical Network Integrated Computational Explorer (BNICE) and expert-curated reaction rules to generate more than 570,000 biofeasible structures starting from 18,000 KEGG metabolites.[122] At the time of writing MS-FINDER integrates structures and formulae for 224,622 known metabolites[94] and also includes 643,307 hypothetical metabolites from MINE-DB.[122] The advantage of these biologically based *in silico* biotransformation methods is that the enzymatic reaction steps and enzymatic mechanisms are explicitly shown or referenced. In other words, the rationale and provenance for each predicted compound is available. The disadvantage is that these biotransformation programs can occasionally produce unreasonable combinatorial explosions. Likewise, they can't make "out-of-the-box" predictions or generate non-obvious or unexpected metabolites. An expansion of this approach was recently published to encompass likely occurring chemical damage (such as oxidations) of molecules, in an analogous database called CD-MINE to cover spontaneously occurring chemical transformations.[123]

Guidelines for Appropriate Reference Library Size and Composition

Both the size and composition of reference libraries will impact the assessment of metabolite identification probability. A reference library that is too small can result in reduced false discovery rate and seemingly accurate, and thus overly confident, identification probabilities. One that is too large can result in increased false discovery due to the addition of compounds that are highly unlikely to be found in such a sample and reduced identification probabilities.[30] Similarly, one should select the appropriate source of compounds to include in the reference library for a given sample type and use case. For example, if a study focuses on a specific organism in a laboratory-controlled setting, then only those molecules potentially produced or consumed by the organism, present in growth media, for example, or known as common contaminants present in the chosen analytical method should be included in the reference library. That is, to prevent misidentifications, one should use organism-specific or sample-specific reference libraries of appropriate size and composition. By comparison, the proteomics community typically uses an appropriate protein FASTA file containing the amino acid sequences of all proteins expected in the organism(s) under study and that are based on translations of the corresponding genomes when searching peptide MS/MS spectra.

Reference libraries for studies of specific organisms in controlled laboratory settings

Different organisms can have profoundly different metabolic needs and metabolic capabilities. For instance, plants have very different metabolomes than animals.[124] Furthermore, the regular consumption of processed foods, supplements, and drugs by humans means that people will have a very different metabolome than lab rats raised on strict chow diet. Indeed, direct comparison of plasma metabolomes showed that less than half the LC-MS signals were common to seven different mammalian species.[125] These results argue for the need for appropriately specialized reference libraries for metabolomics studies to ensure the reliability of metabolite identification.[126] For studies of specific organisms in controlled laboratory settings, we recommend starting with reference libraries that are populated based upon either genome-enabled metabolic reconstructions (e.g., genome-scale metabolic models)[127] or comprehensive review and curation of the literature in respect to metabolomics and other studies of metabolisms of specific organisms. There are a number of existing genome-scale metabolic models for certain organisms, such as *E. coli*,[128] *S. cerevisiae*,[129] *M. musculus*, *D. rerio* and *D. melanogaster*,[130] as well as methods and resources for the scientific community to continue to expand these models or apply them to new organisms.[131, 132] Of note, some of these models overly depend on genomic inference and have rather sparse metabolite information, for which metabolomics can contribute significantly to their expansion.[23] Similarly, there are a number of reference libraries for specific organisms and derived from comprehensive review and curation of the literature, such as the Yeast Metabolome Database[133], the *E. coli* Metabolome Database[134], and the *Pseudomonas aeruginosa* Metabolome Database.[135] PathBank contains literature-derived metabolome reference libraries from several other model organisms.[136] When studying specific organisms in controlled laboratory settings, our recommendation is to use a reference library derived from comprehensive review and curation of the literature. If one is not available, then use an appropriate genome-scale metabolic model. Selecting a library from another organism that is the closest taxonomical relative can also be useful in supplementing a suspect library for an organism that has not been well studied. Finally, the organism-specific metabolite reference libraries should be supplemented with additional inputs from the experiment (e.g., growth media components), including common contaminants present in the chosen analytical method and that are likely to be identified in the experimental data (e.g., plasticizers).

Reference libraries for studies of free-living organisms or environmental systems

As mentioned above, additional considerations are necessary when developing reference libraries for free-living (e.g., humans) or environmental (e.g., soils, forests) systems. Such organisms or systems are not constrained to controlled settings and experience various and diverse inputs to their metabolomes on routine if not daily bases. Further, even within a given free-living system, a human subject for example, the metabolome of one organ or organ system can be very different than that of another. For example, the human blood metabolome is very different from the human urine metabolome.[137, 138] Likewise, the plant leaf metabolome is very different from the associated plant rhizosphere.[139] A number of reference libraries have been developed for free-living organisms. These include the previously mentioned HMDB,[44] the Livestock Metabolome Database,[140] and the Bovine Metabolome Database.[141] Similarly, a variety of matrix-specific or biofluid-specific resources such as the Fecal Metabolome Database,[142] the Saliva Metabolome Database,[143] the Serum Metabolome Database,[137] and the Urine Metabolome Database[138] have also been publicly released. Such databases, or metabolome atlases, may also include aspects of the impact of disease or other factors. As an example, recently the Metabolome Atlas of the Aging Mouse Brain was published.[144] As with laboratory-controlled systems, reference libraries for free-living organisms and environmental systems should be supplemented with other molecules that might be expected to be present in the organism or sample of interest, based on typical behaviors or environmental exposures. Examples of such molecules are contained within the Blood Exposome Database,[145] which covers compounds identified in human blood; DrugBank,[146] which covers approved drugs found in humans; the Toxic Exposome

Database,[147] which covers toxic compounds found in humans; the Norman Suspect List Exchange,[148] which covers common environmental or water contaminants and FooDB (https://foodb.ca), which covers food compounds and food additives found in foods consumed by humans. Further, and as described above, researchers should supplement reference libraries for free-living organisms or environmental systems with information and properties for molecules that are generated via *in silico* predictions of relevant biotransformations or from *in vitro* or cellular incubations.[24] Finally, the comprehensive reference libraries for free-living organisms and environmental systems should be supplemented with additional inputs from the experiment, including common contaminants present in the chosen analytical method and that are likely to be identified in the experimental data.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE IMPLEMENTATION OF METABOLITE IDENTIFICATION PROBABILITY

In this perspective, we have introduced a new concept of metabolite identification probability and have demonstrated its utility in mock identifications using reference libraries constructed from subsets of HMDB and computationally generated RT, CCS, *m/z*, and MS/MS data. The method is computationally simple, automatable, and transferable among analytical platforms. It requires only processed metabolomics data, appropriately defined tolerances allowed in the associated measurement precisions, and reference libraries that are comprehensive and appropriate for the system being queried. We recommend that the metabolomics and related communities (e.g., the non-target analysis community) join us in further exploring the metabolite identification probability approach to more fully reveal its potential and limitations, using real data from real studies and in parallel with their current preferred methods for assessing metabolite identification confidence (e.g., MSI levels), in order to accumulate data on method performance relevant to state-of-the-art. Further extension of these concepts to unidentified features will be required to fully address e.g., unknown chemical hazards of the exposome.[64, 149]

Metabolite identification probability is heavily dependent on the richness of the experimental data being matched to the reference library, the dimensionality and therefore overall resolution of the analytical measurement, the overall measurement precision(s), and the composition and size of the reference library itself. A key requirement for successful implementation of the metabolite identification probability concept is thus the availability of comprehensive and system-appropriate reference libraries. Further research and discussion within the community are needed to determine the repertoire of metabolites and related molecules that should comprise a reference library for a given system, such that metabolite identification probabilities are neither over- nor underestimated. Related, because of the limitation of commercial availability of reference compounds for all system-relevant small molecules, we recommend that the community begin adopting computational approaches for calculating or predicting the associated observable properties, such as spectra, such that reference libraries can be made complete. The accuracy of computationally predicted data should improve with time as methods and technology improve.

Finally, in order that reported metabolite identification probabilities can be transparent, we recommend that individual laboratories version their in-house reference libraries and make them available to the rest of the community as e.g., open mass spectral libraries (OMSL). Besides increasing transparency in calculations of identification probabilities, versioned OMSL and other libraries will be a tremendous resource to the metabolomics research community, as has already been demonstrated by resources such as GNPS[79] and enabled through workflows such as FragHub.[150] As inspiration for how such sharing might be implemented, the metabolomics community can look to the Universal Protein Knowledgebase (UniProtKB)[151] as an example. UniProtKB is a freely accessible database of curated protein sequences that are used, among other purposes, as "reference libraries" for proteomics data searches.

ACKNOWLEDGMENTS

The concept of metabolite identification probability emanated from discussions held among the Compound Identification Development Cores Sub-Committee of the NIH Common Fund Metabolomics Program Phase II. The authors thank Mr. Nathan Johnson from Pacific Northwest National Laboratory (PNNL) for help in designing Figure 1. T.O.M., C.H.C, V.G., A.A., S.T., F.W. S.M.C., J.R.N., M.R.B., M.R.S., and D.S.W. were supported by the National Institutes of Health, National Institute of Environmental Health Sciences (NIEHS) grant U2CES030170 via the Pacific Northwest Advanced Compound Identification Core. A.S.E. was supported by NIEHS grant U2CES030167. D.P.J. and E.T.M. acknowledge support from NIEHS grant U2CES030163. D.S.W. acknowledges additional support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and from the Canada Foundation for Innovation Major Science Initiative (CFI-MSI). G.J.P. acknowledges support from National Cancer Institute grant U01CA235482 and NIEHS grant R35ES028365. O.F. was supported by NIEHS grant U2CES030158 and National Institute of General Medical Sciences grant R01GM0155383. T.O.M. and D.H.R. acknowledge additional support from the PNNL Laboratory Directed Research and Development Program via the *m/q* Initiative.

PNNL is a multi-program national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RLO 1830.

This manuscript was subjected to the U.S. Environmental Protection Agency internal review process. The research results presented do not necessarily reflect the views of the Agency or its policy. Mention of trade names or products does not constitute endorsement or recommendation for use.

REFERENCES

(1) Sauer, U.; Heinemann, M.; Zamboni, N. Genetics. Getting closer to the whole picture. *Science* **2007**, *316* (5824), 550-551. DOI: 10.1126/science.1142502.

(2) Nurse, P.; Hayles, J. The cell in an era of systems biology. *Cell* **2011**, *144* (6), 850-854. DOI: 10.1016/j.cell.2011.02.045.
(3) Westerhoff, H. V.; Palsson, B. O. The evolution of molecular biology into systems biology. *Nat Biotechnol* **2004**, *22* (10), 1249-1252. DOI: 10.1038/nbt1020.
(4) Giani, A. M.; Gallo, G. R.; Gianfranceschi, L.; Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J* **2020**, *18*, 9-19. DOI: 10.1016/j.csbj.2019.11.002.
(5) Heather, J. M.; Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **2016**, *107* (1), 1-8. DOI: 10.1016/j.ygeno.2015.11.003.
(6) Song, Y.; Xu, X.; Wang, W.; Tian, T.; Zhu, Z.; Yang, C. Single cell transcriptomics: moving towards multi-omics. *Analyst* **2019**, *144* (10), 3172-3189. DOI: 10.1039/c8an01852a.
(7) Aebersold, R.; Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **2016**, *537* (7620), 347-355. DOI: 10.1038/nature19949.
(8) Hashimoto, Y.; Greco, T. M.; Cristea, I. M. Contribution of Mass Spectrometry-Based Proteomics to Discoveries in Developmental Biology. *Adv Exp Med Biol* **2019**, *1140*, 143-154. DOI: 10.1007/978-3-030-15950-4_8.
(9) Nicholson, J. K.; Lindon, J. C. Systems biology: Metabonomics. *Nature* **2008**, *455* (7216), 1054-1056. DOI: 10.1038/4551054a.
(10) Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol* **2002**, *48* (1-2), 155-171.
(11) Watson, J. D.; Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **1953**, *171* (4356), 737-738. DOI: 10.1038/171737a0.
(12) Franklin, R. E.; Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **1953**, *171* (4356), 740-741. DOI: 10.1038/171740a0.
(13) Schmutz, J.; Wheeler, J.; Grimwood, J.; Dickson, M.; Yang, J.; Caoile, C.; Bajorek, E.; Black, S.; Chan, Y. M.; Denys, M.; et al. Quality assessment of the human genome sequence. *Nature* **2004**, *429* (6990), 365-368. DOI: 10.1038/nature02390.
(14) Lou, D. I.; Hussmann, J. A.; McBee, R. M.; Acevedo, A.; Andino, R.; Press, W. H.; Sawyer, S. L. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci U S A* **2013**, *110* (49), 19872-19877. DOI: 10.1073/pnas.1319590110 From NLM Medline.
(15) Mewes, H. W.; Amid, C.; Arnold, R.; Frishman, D.; Guldener, U.; Mannhaupt, G.; Munsterkotter, M.; Pagel, P.; Strack, N.; Stumpflen, V.; et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **2004**, *32* (Database issue), D41-44. DOI: 10.1093/nar/gkh092.
(16) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **1994**, *5* (11), 976-989. DOI: 10.1016/1044-0305(94)80016-2.
(17) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551-3567. DOI: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.
(18) Ma, K.; Vitek, O.; Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* **2012**, *13 Suppl 16*, S1. DOI: 10.1186/1471-2105-13-S16-S1.
(19) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007**, *4* (3), 207-214. DOI: 10.1038/nmeth1019.
(20) Dunphy, K.; Dowling, P.; Bazou, D.; O'Gorman, P. Current Methods of Post-Translational Modification Analysis and Their Applications in Blood Cancers. *Cancers (Basel)* **2021**, *13* (8). DOI: 10.3390/cancers13081930.

(21) Delong, T.; Wiles, T. A.; Baker, R. L.; Bradley, B.; Barbour, G.; Reisdorph, R.; Armstrong, M.; Powell, R. L.; Reisdorph, N.; Kumar, N.; et al. Pathogenic CD4 T cells in type 1 diabetes recognize epitopes formed by peptide fusion. *Science* **2016**, *351* (6274), 711-714. DOI: 10.1126/science.aad2791.

(22) Montenegro-Burke, J. R.; Guijas, C.; Siuzdak, G. METLIN: A Tandem Mass Spectral Library of Standards. *Methods Mol Biol* **2020**, *2104*, 149-163. DOI: 10.1007/978-1-0716-0239-3_9.

(23) Frainay, C.; Schymanski, E. L.; Neumann, S.; Merlet, B.; Salek, R. M.; Jourdan, F.; Yanes, O. Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas. *Metabolites* **2018**, *8* (3). DOI: 10.3390/metabo8030051.

(24) Liu, K. H.; Lee, C. M.; Singer, G.; Bais, P.; Castellanos, F.; Woodworth, M. H.; Ziegler, T. R.; Kraft, C. S.; Miller, G. W.; Li, S.; et al. Large scale enzyme based xenobiotic identification for exposomics. *Nat Commun* **2021**, *12* (1), 5418. DOI: 10.1038/s41467-021-25698-x.

(25) Stanstrup, J.; Broeckling, C. D.; Helmus, R.; Hoffmann, N.; Mathe, E.; Naake, T.; Nicolotti, L.; Peters, K.; Rainer, J.; Salek, R. M.; et al. The metaRbolomics Toolbox in Bioconductor and beyond. *Metabolites* **2019**, *9* (10). DOI: 10.3390/metabo9100200.

(26) Mitchell, J. M.; Chi, Y.; Thapa, M.; Pang, Z.; Xia, J.; Li, S. Common data models to streamline metabolomics processing and annotation, and implementation in a Python pipeline. *PLoS Comput Biol* **2024**, *20* (6), e1011912. DOI: 10.1371/journal.pcbi.1011912 From NLM Medline.

(27) Stein, S. E. Estimating probabilities of correct identification from results of mass spectral library searches. *J Am Soc Mass Spectrom* **1994**, *5* (4), 316-323. DOI: 10.1016/1044-0305(94)85022-4.

(28) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007**, *3* (3), 211-221. DOI: 10.1007/s11306-007-0082-2.

(29) Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat Methods* **2021**. DOI: 10.1038/s41592-021-01331-z.

(30) Matsuda, F.; Shinbo, Y.; Oikawa, A.; Hirai, M. Y.; Fiehn, O.; Kanaya, S.; Saito, K. Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches. *PLoS One* **2009**, *4* (10), e7490. DOI: 10.1371/journal.pone.0007490.

(31) Scheubert, K.; Hufsky, F.; Petras, D.; Wang, M.; Nothias, L. F.; Duhrkop, K.; Bandeira, N.; Dorrestein, P. C.; Bocker, S. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun* **2017**, *8* (1), 1494. DOI: 10.1038/s41467-017-01318-5.

(32) Castle, A. L.; Fiehn, O.; Kaddurah-Daouk, R.; Lindon, J. C. Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. *Brief Bioinform* **2006**, *7* (2), 159-165. DOI: 10.1093/bib/bbl008.

(33) Fiehn, O.; Kristal, B.; van Ommen, B.; Sumner, L. W.; Sansone, S. A.; Taylor, C.; Hardy, N.; Kaddurah-Daouk, R. Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. *OMICS* **2006**, *10* (2), 158-163. DOI: 10.1089/omi.2006.10.158.

(34) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014**, *48* (4), 2097-2098. DOI: 10.1021/es5002105.

(35) Celma, A.; Sancho, J. V.; Schymanski, E. L.; Fabregat-Safont, D.; Ibanez, M.; Goshawk, J.; Barknowitz, G.; Hernandez, F.; Bijlsma, L. Improving Target and Suspect Screening High-Resolution Mass Spectrometry Workflows in Environmental Analysis by Ion Mobility Separation. *Environ Sci Technol* **2020**, *54* (23), 15120-15131. DOI: 10.1021/acs.est.0c05713.

(36) Sumner, L. W.; Lei, Z.; Nikolau, B. J.; Saito, K.; Roessner, U.; Trengove, R. Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics* **2014**, *10*, 1047-1049.
(37) Creek, D. J.; Dunn, W. B.; Fiehn, O.; Griffin, J. L.; Hall, R. D.; Lei, Z.; Mistrik, R.; Neumann, S.; Schymanski, E. L.; Sumner, L. W.; et al. Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics* **2014**, *10*, 350-353.
(38) Alygizakis, N.; Lestremau, F.; Gago-Ferrero, P.; Gil-Solsona, R.; Arturi, K.; Hollender, J.; Schymanski, E. L.; Dulio, V.; Slobodnik, J.; Thomaidis, N. S. Towards a harmonized identification scoring system in LC-HRMS/MS based non-target screening (NTS) of emerging contaminants. *Trac-Trend Anal Chem* **2023**, *159*. DOI: ARTN 116944 10.1016/j.trac.2023.116944.
(39) 2002/657/EC: Commission Decision of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results (Text with EEA relevance) (notified under document number C(2002) 3044). 2002; pp 8-36.
(40) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J Cheminform* **2013**, *5* (1), 7. DOI: 10.1186/1758-2946-5-7.
(41) Kofeler, H. C.; Eichmann, T. O.; Ahrends, R.; Bowden, J. A.; Danne-Rasche, N.; Dennis, E. A.; Fedorova, M.; Griffiths, W. J.; Han, X.; Hartler, J.; et al. Quality control requirements for the correct annotation of lipidomics data. *Nat Commun* **2021**, *12* (1), 4771. DOI: 10.1038/s41467-021-24984-y.
(42) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* **2016**, *44* (D1), D463-470. DOI: 10.1093/nar/gkv1042.
(43) Haug, K.; Cochrane, K.; Nainala, V. C.; Williams, M.; Chang, J.; Jayaseelan, K. V.; O'Donovan, C. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res* **2020**, *48* (D1), D440-D444. DOI: 10.1093/nar/gkz1019.
(44) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; et al. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res* **2022**, *50* (D1), D622-D631. DOI: 10.1093/nar/gkab1062.
(45) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminform* **2016**, *8*, 61. DOI: 10.1186/s13321-016-0174-y From NLM PubMed-not-MEDLINE.
(46) Colby, S. M.; Nunez, J. R.; Hodas, N. O.; Corley, C. D.; Renslow, R. R. Deep Learning to Generate in Silico Chemical Property Libraries and Candidate Molecules for Small Molecule Identification in Complex Samples. *Anal Chem* **2020**, *92* (2), 1720-1729. DOI: 10.1021/acs.analchem.9b02348.
(47) Zhou, Z.; Luo, M.; Chen, X.; Yin, Y.; Xiong, X.; Wang, R.; Zhu, Z. J. Ion mobility collision cross-section atlas for known and unknown metabolite annotation in untargeted metabolomics. *Nat Commun* **2020**, *11* (1), 4334. DOI: 10.1038/s41467-020-18171-8.
(48) Plante, P. L.; Francovic-Fontaine, E.; May, J. C.; McLean, J. A.; Baker, E. S.; Laviolette, F.; Marchand, M.; Corbeil, J. Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS. *Anal Chem* **2019**, *91* (8), 5191-5199. DOI: 10.1021/acs.analchem.8b05821.
(49) Bonini, P.; Kind, T.; Tsugawa, H.; Barupal, D. K.; Fiehn, O. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal Chem* **2020**, *92* (11), 7515-7522. DOI: 10.1021/acs.analchem.9b05765.

(50) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal Chem* **2021**, *93* (34), 11692-11700. DOI: 10.1021/acs.analchem.1c01465 From NLM Medline.
(51) Wojcik, R.; Nagy, G.; Attah, I. K.; Webb, I. K.; Garimella, S. V. B.; Weitz, K. K.; Hollerbach, A.; Monroe, M. E.; Ligare, M. R.; Nielson, F. F.; et al. SLIM Ultrahigh Resolution Ion Mobility Spectrometry Separations of Isotopologues and Isotopomers Reveal Mobility Shifts due to Mass Distribution Changes. *Anal Chem* **2019**, *91* (18), 11952-11962. DOI: 10.1021/acs.analchem.9b02808 From NLM Medline.
(52) Rout, M.; Lipfert, M.; Lee, B. L.; Berjanskii, M.; Assempour, N.; Fresno, R. V.; Cayuela, A. S.; Dong, Y.; Johnson, M.; Shahin, H.; et al. MagMet: A fully automated web server for targeted nuclear magnetic resonance metabolomics of plasma and serum. *Magn Reson Chem* **2023**, *61* (12), 681-704. DOI: 10.1002/mrc.5371 From NLM Medline.
(53) Ravanbakhsh, S.; Liu, P.; Bjorndahl, T. C.; Mandal, R.; Grant, J. R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; et al. Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS One* **2015**, *10* (5), e0124219. DOI: 10.1371/journal.pone.0124219 From NLM Medline.
(54) Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal Chem* **2006**, *78* (13), 4430-4442. DOI: 10.1021/ac060209g From NLM Medline.
(55) Kind, T.; Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **2007**, *8*, 105. DOI: 10.1186/1471-2105-8-105.
(56) Ludwig, M.; Fleischauer, M.; Duhrkop, K.; Hoffmann, M. A.; Bocker, S. De Novo Molecular Formula Annotation and Structure Elucidation Using SIRIUS 4. *Methods Mol Biol* **2020**, *2104*, 185-207. DOI: 10.1007/978-1-0716-0239-3_11.
(57) Pluskal, T.; Uehara, T.; Yanagida, M. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal Chem* **2012**, *84* (10), 4396-4403. DOI: 10.1021/ac3000418.
(58) Duhrkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Bocker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A* **2015**, *112* (41), 12580-12585. DOI: 10.1073/pnas.1509788112.
(59) Draper, J.; Enot, D. P.; Parker, D.; Beckmann, M.; Snowdon, S.; Lin, W.; Zubair, H. Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinformatics* **2009**, *10*, 227. DOI: 10.1186/1471-2105-10-227.
(60) A, G. M.; G, T. B.; Chen, T.; N, K. K.; A, M. M.; R, P. R.; B, M. R.; Xian, F. Mass resolution and mass accuracy: how much is enough? *Mass Spectrom (Tokyo)* **2013**, *2* (Spec Iss), S0009. DOI: 10.5702/massspectrometry.S0009 From NLM PubMed-not-MEDLINE.
(61) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* **2021**, *49* (D1), D1388-D1395. DOI: 10.1093/nar/gkaa971.
(62) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012**, *40* (Database issue), D1100-1107. DOI: 10.1093/nar/gkr777.
(63) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform* **2017**, *9* (1), 61. DOI: 10.1186/s13321-017-0247-6.

(64) Vermeulen, R.; Schymanski, E. L.; Barabasi, A. L.; Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **2020**, *367* (6476), 392-396. DOI: 10.1126/science.aay3164 From NLM Medline.
(65) Sud, M.; Fahy, E.; Cotter, D.; Dennis, E. A.; Subramaniam, S. LIPID MAPS-Nature Lipidomics Gateway: An Online Resource for Students and Educators Interested in Lipids. *J Chem Educ* **2012**, *89* (2), 291-292. DOI: 10.1021/ed200088u.
(66) Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000**, *28* (1), 27-30. DOI: 10.1093/nar/28.1.27.
(67) Caspi, R.; Billington, R.; Fulcher, C. A.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Midford, P. E.; Ong, Q.; Ong, W. K.; et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* **2018**, *46* (D1), D633-D639. DOI: 10.1093/nar/gkx935.
(68) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **2010**, *87* (11), 1123-1124. DOI: 10.1021/ed100697w.
(69) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem Substance and Compound databases. *Nucleic Acids Res* **2016**, *44* (D1), D1202-1213. DOI: 10.1093/nar/gkv951 From NLM Medline.
(70) Braisted, J.; Patt, A.; Tindall, C.; Sheils, T.; Neyra, J.; Spencer, K.; Eicher, T.; Mathe, E. A. RaMP-DB 2.0: a renovated knowledgebase for deriving biological and chemical insight from metabolites, proteins, and genes. *Bioinformatics* **2023**, *39* (1). DOI: 10.1093/bioinformatics/btac726 From NLM Medline.
(71) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **2016**, *44* (D1), D1214-1219. DOI: 10.1093/nar/gkv1031 From NLM Medline.
(72) van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castano-Espriu, L.; Chang, C.; Clark, T. N.; et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent Sci* **2019**, *5* (11), 1824-1833. DOI: 10.1021/acscentsci.9b00806 From NLM PubMed-not-MEDLINE.
(73) Kind, T.; Fiehn, O. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* **2006**, *7*, 234. DOI: 10.1186/1471-2105-7-234.
(74) Witting, M.; Bocker, S. Current status of retention time prediction in metabolite identification. *J Sep Sci* **2020**, *43* (9-10), 1746-1754. DOI: 10.1002/jssc.202000060.
(75) Khanal, N.; Masellis, C.; Kamrath, M. Z.; Clemmer, D. E.; Rizzo, T. R. Cryogenic IR spectroscopy combined with ion mobility spectrometry for the analysis of human milk oligosaccharides. *Analyst* **2018**, *143* (8), 1846-1852. DOI: 10.1039/c8an00230d From NLM Medline.
(76) Huber, F.; Ridder, L.; Verhoeven, S.; Spaaks, J. H.; Diblen, F.; Rogers, S.; van der Hooft, J. J. J. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput Biol* **2021**, *17* (2), e1008724. DOI: 10.1371/journal.pcbi.1008724.
(77) Stein, S. E.; Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom* **1994**, *5* (9), 859-866. DOI: 10.1016/1044-0305(94)87009-8.
(78) Stow, S. M.; Causon, T. J.; Zheng, X.; Kurulugama, R. T.; Mairinger, T.; May, J. C.; Rennie, E. E.; Baker, E. S.; Smith, R. D.; McLean, J. A.; et al. An Interlaboratory Evaluation of Drift Tube Ion Mobility-Mass Spectrometry Collision Cross Section Measurements. *Anal Chem* **2017**, *89* (17), 9048-9055. DOI: 10.1021/acs.analchem.7b01729.
(79) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; et al. Sharing and community curation of mass

spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **2016**, *34* (8), 828-837. DOI: 10.1038/nbt.3597.

(80) Romero, P. R.; Kobayashi, N.; Wedell, J. R.; Baskaran, K.; Iwata, T.; Yokochi, M.; Maziuk, D.; Yao, H.; Fujiwara, T.; Kurusu, G.; et al. BioMagResBank (BMRB) as a Resource for Structural Biology. *Methods Mol Biol* **2020**, *2112*, 187-218. DOI: 10.1007/978-1-0716-0270-6_14.

(81) Steinbeck, C.; Kuhn, S. NMRShiftDB -- compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **2004**, *65* (19), 2711-2717. DOI: 10.1016/j.phytochem.2004.08.027.

(82) Wishart, D. S.; Sayeeda, Z.; Budinski, Z.; Guo, A.; Lee, B. L.; Berjanskii, M.; Rout, M.; Peters, H.; Dizon, R.; Mah, R.; et al. NP-MRD: the Natural Products Magnetic Resonance Database. *Nucleic Acids Res* **2021**. DOI: 10.1093/nar/gkab1052.

(83) Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R. Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server. *Anal Chem* **2016**, *88* (24), 12411-12418. DOI: 10.1021/acs.analchem.6b03724 From NLM Medline.

(84) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal Chem* **2009**, *81* (24), 10038-10048. DOI: 10.1021/ac9019522.

(85) Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Collision cross section compendium to annotate and predict multi-omic compound identities. *Chem Sci* **2019**, *10* (4), 983-993. DOI: 10.1039/c8sc04396e.

(86) Schroeder, M.; Meyer, S. W.; Heyman, H. M.; Barsch, A.; Sumner, L. W. Generation of a Collision Cross Section Library for Multi-Dimensional Plant Metabolomics Using UHPLC-Trapped Ion Mobility-MS/MS. *Metabolites* **2019**, *10* (1). DOI: 10.3390/metabo10010013.

(87) Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J. R.; Uritboonthai, W.; Aisporna, A. E.; Chen, E.; Benton, H. P.; Siuzdak, G. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nat Commun* **2019**, *10* (1), 5811. DOI: 10.1038/s41467-019-13680-7.

(88) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. METLIN: a metabolite mass spectral database. *Ther Drug Monit* **2005**, *27* (6), 747-751. DOI: 10.1097/01.ftd.0000179845.53213.39 From NLM Medline.

(89) Gauglitz, J. M.; West, K. A.; Bittremieux, W.; Williams, C. L.; Weldon, K. C.; Panitchpakdi, M.; Di Ottavio, F.; Aceves, C. M.; Brown, E.; Sikora, N. C.; et al. Enhancing untargeted metabolomics using metadata-based source annotation. *Nat Biotechnol* **2022**, *40* (12), 1774-1779. DOI: 10.1038/s41587-022-01368-1 From NLM Medline.

(90) Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods* **2018**, *15* (1), 53-56. DOI: 10.1038/nmeth.4512 From NLM Medline.

(91) Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz, K. M., Jr.; Metz, T. O.; et al. Quantum Chemistry Calculations for Metabolomics. *Chem Rev* **2021**, *121* (10), 5633-5670. DOI: 10.1021/acs.chemrev.0c00901.

(92) Wolf, S.; Schmidt, S.; Muller-Hannemann, M.; Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **2010**, *11*, 148. DOI: 10.1186/1471-2105-11-148.

(93) Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* **2014**, *42* (Web Server issue), W94-99. DOI: 10.1093/nar/gku436.

(94) Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal Chem* **2016**, *88* (16), 7946-7958. DOI: 10.1021/acs.analchem.6b00770.
(95) Laponogov, I.; Sadawi, N.; Galea, D.; Mirnezami, R.; Veselkov, K. A. ChemDistiller: an engine for metabolite annotation in mass spectrometry. *Bioinformatics* **2018**, *34* (12), 2096-2102. DOI: 10.1093/bioinformatics/bty080.
(96) Ridder, L.; van der Hooft, J. J.; Verhoeven, S. Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrom (Tokyo)* **2014**, *3* (Spec Iss 2), S0033. DOI: 10.5702/massspectrometry.S0033.
(97) Schymanski, E. L.; Ruttkies, C.; Krauss, M.; Brouard, C.; Kind, T.; Duhrkop, K.; Allen, F.; Vaniya, A.; Verdegem, D.; Bocker, S.; et al. Critical Assessment of Small Molecule Identification 2016: automated methods. *J Cheminform* **2017**, *9* (1), 22. DOI: 10.1186/s13321-017-0207-1.
(98) Chao, A.; Al-Ghoul, H.; McEachran, A. D.; Balabin, I.; Transue, T.; Cathey, T.; Grossman, J. N.; Singh, R. R.; Ulrich, E. M.; Williams, A. J.; et al. In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples. *Anal Bioanal Chem* **2020**, *412* (6), 1303-1315. DOI: 10.1007/s00216-019-02351-7.
(99) Allen, F.; Pon, A.; Greiner, R.; Wishart, D. Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification. *Anal Chem* **2016**, *88* (15), 7689-7697. DOI: 10.1021/acs.analchem.6b01622.
(100) Ji, H.; Deng, H.; Lu, H.; Zhang, Z. Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks. *Anal Chem* **2020**, *92* (13), 8649-8653. DOI: 10.1021/acs.analchem.0c01450.
(101) Asgeirsson, V.; Bauer, C. A.; Grimme, S. Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules. *Chem Sci* **2017**, *8* (7), 4879-4895. DOI: 10.1039/c7sc00601b.
(102) Koopman, J.; Grimme, S. From QCEIMS to QCxMS: A Tool to Routinely Calculate CID Mass Spectra Using Molecular Dynamics. *J Am Soc Mass Spectrom* **2021**, *32* (7), 1735-1751. DOI: 10.1021/jasms.1c00098 From NLM PubMed-not-MEDLINE.
(103) Das, S.; Edison, A. S.; Merz, K. M., Jr. Metabolite Structure Assignment Using In Silico NMR Techniques. *Anal Chem* **2020**, *92* (15), 10412-10419. DOI: 10.1021/acs.analchem.0c00768.
(104) Hoffmann, F.; Li, D. W.; Sebastiani, D.; Bruschweiler, R. Improved Quantum Chemical NMR Chemical Shift Prediction of Metabolites in Aqueous Solution toward the Validation of Unknowns. *J Phys Chem A* **2017**, *121* (16), 3071-3078. DOI: 10.1021/acs.jpca.7b01954.
(105) Wang, B.; Dossey, A. T.; Walse, S. S.; Edison, A. S.; Merz, K. M., Jr. Relative configuration of natural products using NMR chemical shifts. *J Nat Prod* **2009**, *72* (4), 709-713. DOI: 10.1021/np8005056 From NLM Medline.
(106) Yesiltepe, Y.; Nunez, J. R.; Colby, S. M.; Thomas, D. G.; Borkum, M. I.; Reardon, P. N.; Washton, N. M.; Metz, T. O.; Teeguarden, J. G.; Govind, N.; et al. An automated framework for NMR chemical shift calculations of small organic molecules. *J Cheminform* **2018**, *10* (1), 52. DOI: 10.1186/s13321-018-0305-8.
(107) Gao, P.; Zhang, J.; Peng, Q.; Zhang, J.; Glezakou, V. A. General Protocol for the Accurate Prediction of Molecular (13)C/(1)H NMR Chemical Shifts via Machine Learning Augmented DFT. *J Chem Inf Model* **2020**, *60* (8), 3746-3754. DOI: 10.1021/acs.jcim.0c00388.
(108) Sajed, T.; Sayeeda, Z.; Lee, B. L.; Berjanskii, M.; Wang, F.; Gautam, V.; Wishart, D. S. Accurate Prediction of (1)H NMR Chemical Shifts of Small Molecules Using Machine Learning. *Metabolites* **2024**, *14* (5). DOI: 10.3390/metabo14050290 From NLM PubMed-not-MEDLINE.
(109) Zhou, Z.; Xiong, X.; Zhu, Z. J. MetCCS predictor: a web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. *Bioinformatics* **2017**, *33* (14), 2235-2237. DOI: 10.1093/bioinformatics/btx140.

(110) Colby, S. M.; Thomas, D. G.; Nunez, J. R.; Baxter, D. J.; Glaesemann, K. R.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J. G.; Metz, T. O.; et al. ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries. *Anal Chem* **2019**, *91* (7), 4346-4356. DOI: 10.1021/acs.analchem.8b04567.
(111) Ross, D. H.; Cho, J. H.; Xu, L. Breaking Down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections. *Anal Chem* **2020**, *92* (6), 4548-4557. DOI: 10.1021/acs.analchem.9b05772.
(112) Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z. J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Anal Chem* **2016**, *88* (22), 11084-11091. DOI: 10.1021/acs.analchem.6b03091.
(113) Qu, C.; Schneider, B. I.; Kearsley, A. J.; Keyrouz, W.; Allison, T. C. Predicting Kovats Retention Indices Using Graph Neural Networks. *J Chromatogr A* **2021**, *1646*, 462100. DOI: 10.1016/j.chroma.2021.462100.
(114) Yang, Q.; Ji, H.; Lu, H.; Zhang, Z. Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification. *Anal Chem* **2021**, *93* (4), 2200-2206. DOI: 10.1021/acs.analchem.0c04071.
(115) Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminform* **2019**, *11* (1), 2. DOI: 10.1186/s13321-018-0324-5.
(116) Bian, Y.; Xie, X. Q. Generative chemistry: drug discovery with deep learning generative models. *J Mol Model* **2021**, *27* (3), 71. DOI: 10.1007/s00894-021-04674-8.
(117) Long, A. Drug metabolism in silico - the knowledge-based expert system approach. Historical perspectives and current strategies. *Drug Discov Today Technol* **2013**, *10* (1), e147-153. DOI: 10.1016/j.ddtec.2012.10.006.
(118) de Bruyn Kops, C.; Sicho, M.; Mazzolari, A.; Kirchmair, J. GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics. *Chem Res Toxicol* **2021**, *34* (2), 286-299. DOI: 10.1021/acs.chemrestox.0c00224.
(119) Sicho, M.; de Bruyn Kops, C.; Stork, C.; Svozil, D.; Kirchmair, J. FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity. *J Chem Inf Model* **2017**, *57* (8), 1832-1846. DOI: 10.1021/acs.jcim.7b00250.
(120) Sicho, M.; Stork, C.; Mazzolari, A.; de Bruyn Kops, C.; Pedretti, A.; Testa, B.; Vistoli, G.; Svozil, D.; Kirchmair, J. FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes. *J Chem Inf Model* **2019**, *59* (8), 3400-3412. DOI: 10.1021/acs.jcim.9b00376.
(121) Tian, S.; Cao, X.; Greiner, R.; Li, C.; Guo, A.; Wishart, D. S. CyProduct: A Software Tool for Accurately Predicting the Byproducts of Human Cytochrome P450 Metabolism. *J Chem Inf Model* **2021**, *61* (6), 3128-3140. DOI: 10.1021/acs.jcim.1c00144.
(122) Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt, L. J.; Hanson, A. D.; Fiehn, O.; Tyo, K. E.; Henry, C. S. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J Cheminform* **2015**, *7*, 44. DOI: 10.1186/s13321-015-0087-1.
(123) Jeffryes, J. G.; Lerma-Ortiz, C.; Liu, F.; Golubev, A.; Niehaus, T. D.; Elbadawi-Sidhu, M.; Fiehn, O.; Hanson, A. D.; Tyo, K. E.; Henry, C. S. Chemical-damage MINE: A database of curated and predicted spontaneous metabolic reactions. *Metab Eng* **2022**, *69*, 302-312. DOI: 10.1016/j.ymben.2021.11.009.
(124) Edison, A. S.; Hall, R. D.; Junot, C.; Karp, P. D.; Kurland, I. J.; Mistrik, R.; Reed, L. K.; Saito, K.; Salek, R. M.; Steinbeck, C.; et al. The Time Is Right to Focus on Model Organism Metabolomes. *Metabolites* **2016**, *6* (1). DOI: 10.3390/metabo6010008.
(125) Park, Y. H.; Lee, K.; Soltow, Q. A.; Strobel, F. H.; Brigham, K. L.; Parker, R. E.; Wilson, M. E.; Sutliff, R. L.; Mansfield, K. G.; Wachtman, L. M.; et al. High-performance metabolic profiling

of plasma from seven mammalian species for simultaneous environmental chemical surveillance and bioeffect monitoring. *Toxicology* **2012**, *295* (1-3), 47-55. DOI: 10.1016/j.tox.2012.02.007 From NLM Medline.
(126) Rutz, A.; Dounoue-Kubo, M.; Ollivier, S.; Bisson, J.; Bagheri, M.; Saesong, T.; Ebrahimi, S. N.; Ingkaninan, K.; Wolfender, J. L.; Allard, P. M. Taxonomically Informed Scoring Enhances Confidence in Natural Products Annotation. *Front Plant Sci* **2019**, *10*, 1329. DOI: 10.3389/fpls.2019.01329.
(127) Bordbar, A.; Monk, J. M.; King, Z. A.; Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* **2014**, *15* (2), 107-120. DOI: 10.1038/nrg3643.
(128) Edwards, J. S.; Palsson, B. O. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A* **2000**, *97* (10), 5528-5533. DOI: 10.1073/pnas.97.10.5528.
(129) Oftadeh, O.; Salvy, P.; Masid, M.; Curvat, M.; Miskovic, L.; Hatzimanikatis, V. A genome-scale metabolic model of Saccharomyces cerevisiae that integrates expression constraints and reaction thermodynamics. *Nat Commun* **2021**, *12* (1), 4790. DOI: 10.1038/s41467-021-25158-6.
(130) Wang, H.; Robinson, J. L.; Kocabas, P.; Gustafsson, J.; Anton, M.; Cholley, P. E.; Huang, S.; Gobom, J.; Svensson, T.; Uhlen, M.; et al. Genome-scale metabolic network reconstruction of model animals as a platform for translational research. *Proc Natl Acad Sci U S A* **2021**, *118* (30). DOI: 10.1073/pnas.2102344118.
(131) Mendoza, S. N.; Olivier, B. G.; Molenaar, D.; Teusink, B. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol* **2019**, *20* (1), 158. DOI: 10.1186/s13059-019-1769-1.
(132) Macklin, D. N.; Ahn-Horst, T. A.; Choi, H.; Ruggero, N. A.; Carrera, J.; Mason, J. C.; Sun, G.; Agmon, E.; DeFelice, M. M.; Maayan, I.; et al. Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. *Science* **2020**, *369* (6502). DOI: 10.1126/science.aav3751 From NLM Medline.
(133) Ramirez-Gaona, M.; Marcu, A.; Pon, A.; Guo, A. C.; Sajed, T.; Wishart, N. A.; Karu, N.; Djoumbou Feunang, Y.; Arndt, D.; Wishart, D. S. YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res* **2017**, *45* (D1), D440-D445. DOI: 10.1093/nar/gkw1058.
(134) Sajed, T.; Marcu, A.; Ramirez, M.; Pon, A.; Guo, A. C.; Knox, C.; Wilson, M.; Grant, J. R.; Djoumbou, Y.; Wishart, D. S. ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Res* **2016**, *44* (D1), D495-501. DOI: 10.1093/nar/gkv1060.
(135) Huang, W.; Brewer, L. K.; Jones, J. W.; Nguyen, A. T.; Marcu, A.; Wishart, D. S.; Oglesby-Sherrouse, A. G.; Kane, M. A.; Wilks, A. PAMDB: a comprehensive Pseudomonas aeruginosa metabolome database. *Nucleic Acids Res* **2018**, *46* (D1), D575-D580. DOI: 10.1093/nar/gkx1061.
(136) Wishart, D. S.; Li, C.; Marcu, A.; Badran, H.; Pon, A.; Budinski, Z.; Patron, J.; Lipton, D.; Cao, X.; Oler, E.; et al. PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Res* **2020**, *48* (D1), D470-D478. DOI: 10.1093/nar/gkz861.
(137) Psychogios, N.; Hau, D. D.; Peng, J.; Guo, A. C.; Mandal, R.; Bouatra, S.; Sinelnikov, I.; Krishnamurthy, R.; Eisner, R.; Gautam, B.; et al. The human serum metabolome. *PLoS One* **2011**, *6* (2), e16957. DOI: 10.1371/journal.pone.0016957.
(138) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; et al. The human urine metabolome. *PLoS One* **2013**, *8* (9), e73076. DOI: 10.1371/journal.pone.0073076.
(139) Mashabela, M. D.; Tugizimana, F.; Steenkamp, P. A.; Piater, L. A.; Dubery, I. A.; Mhlongo, M. I. Untargeted metabolite profiling to elucidate rhizosphere and leaf metabolome changes of wheat cultivars (Triticum aestivum L.) treated with the plant growth-promoting rhizobacteria

Paenibacillus alvei (T22) and Bacillus subtilis. *Front Microbiol* **2022**, *13*, 971836. DOI: 10.3389/fmicb.2022.971836 From NLM PubMed-not-MEDLINE.
(140) Goldansaz, S. A.; Guo, A. C.; Sajed, T.; Steele, M. A.; Plastow, G. S.; Wishart, D. S. Livestock metabolomics and the livestock metabolome: A systematic review. *PLoS One* **2017**, *12* (5), e0177675. DOI: 10.1371/journal.pone.0177675.
(141) Foroutan, A.; Fitzsimmons, C.; Mandal, R.; Piri-Moghadam, H.; Zheng, J.; Guo, A.; Li, C.; Guan, L. L.; Wishart, D. S. The Bovine Metabolome. *Metabolites* **2020**, *10* (6). DOI: 10.3390/metabo10060233.
(142) Karu, N.; Deng, L.; Slae, M.; Guo, A. C.; Sajed, T.; Huynh, H.; Wine, E.; Wishart, D. S. A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. *Anal Chim Acta* **2018**, *1030*, 1-24. DOI: 10.1016/j.aca.2018.05.031.
(143) Dame, Z. T.; Aziat, F.; Mandal, R.; Krishnamurthy, R.; Bouatra, S.; Borzouie, S.; Guo, A. C.; Sajed, T.; Deng, L.; Lin, H.; et al. The human saliva metabolome. *Metabolomics* **2015**, *11* (6), 1864-1883. DOI: 10.1007/s11306-015-0840-5.
(144) Ding, J.; Ji, J.; Rabow, Z.; Shen, T.; Folz, J.; Brydges, C. R.; Fan, S.; Lu, X.; Mehta, S.; Showalter, M. R.; et al. A metabolome atlas of the aging mouse brain. *Nat Commun* **2021**, *12* (1), 6021. DOI: 10.1038/s41467-021-26310-y.
(145) Barupal, D. K.; Fiehn, O. Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach. *Environ Health Perspect* **2019**, *127* (9), 97008. DOI: 10.1289/EHP4713.
(146) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **2018**, *46* (D1), D1074-D1082. DOI: 10.1093/nar/gkx1037.
(147) Wishart, D.; Arndt, D.; Pon, A.; Sajed, T.; Guo, A. C.; Djoumbou, Y.; Knox, C.; Wilson, M.; Liang, Y.; Grant, J.; et al. T3DB: the toxic exposome database. *Nucleic Acids Res* **2015**, *43* (Database issue), D928-934. DOI: 10.1093/nar/gku1004.
(148) Mohammed Taha, H.; Aalizadeh, R.; Alygizakis, N.; Antignac, J. P.; Arp, H. P. H.; Bade, R.; Baker, N.; Belova, L.; Bijlsma, L.; Bolton, E. E.; et al. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. *Environ Sci Eur* **2022**, *34* (1), 104. DOI: 10.1186/s12302-022-00680-6 From NLM PubMed-not-MEDLINE.
(149) Uppal, K.; Walker, D. I.; Liu, K.; Li, S.; Go, Y. M.; Jones, D. P. Computational Metabolomics: A Framework for the Million Metabolome. *Chem Res Toxicol* **2016**, *29* (12), 1956-1975. DOI: 10.1021/acs.chemrestox.6b00179 From NLM Medline.
(150) Dablanc, A.; Hennechart, S.; Perez, A.; Cabanac, G.; Guitton, Y.; Paulhe, N.; Lyan, B.; Jamin, E. L.; Giacomoni, F.; Marti, G. FragHub: A Mass Spectral Library Data Integration Workflow. *Anal Chem* **2024**. DOI: 10.1021/acs.analchem.4c02219 From NLM Publisher.
(151) UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **2021**, *49* (D1), D480-D489. DOI: 10.1093/nar/gkaa1100.