Title

Introducing “Identification Probability” for Automated and Transferable Assessment of Metabolite Identification Confidence in Metabolomics and Related Studies

Journal

Analytical Chemistry, 97(1)

ISSN

0003-2700

Authors

Metz, Thomas O

Chang, Christine H

Gautam, Vasuk

et al.

Publication Date

2025-01-14

DOI

10.1021/acs.analchem.4c04060

Peer reviewed

analytical chemistry

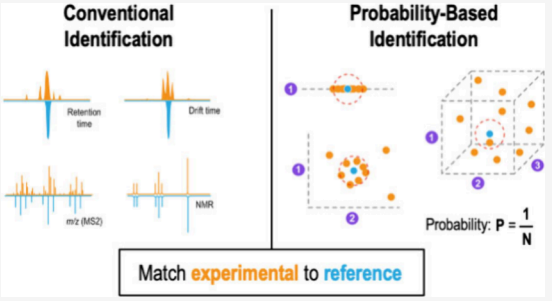
pubs.acs.org/ac
Perspective

Introducing "Identification Probability" for Automated and Transferable Assessment of Metabolite Identification Confidence in Metabolomics and Related Studies

Thomas O. Metz,* Christine H. Chang, Vasuk Gautam, Afia Anjum, Siyang Tian, Fei Wang, Sean M. Colby, Jamie R. Nunez, Madison R. Blumer, Arthur S. Edison, Oliver Fiehn, Dean P. Jones, Shuzhao Li, Edward T. Morgan, Gary J. Patti, Dylan H. Ross, Madelyn R. Shapiro, Antony J. Williams, and David S. Wishart

Cite This: *Anal. Chem.* 2025, 97, 1−11

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 the context of the chemical space being considered. Neither are they easily automated nor 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 database that matches 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 multiproperty reference libraries constructed from a subset of 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

Received: August 1, 2024
Revised: December 2, 2024
Accepted: December 6, 2024
Published: December 19, 2024

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11

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 nonprotein, small molecules (e.g., 50−1500 Da) as possible. Given their high sensitivity and broad molecular coverage, a variety of mass spectrometry-based techniques, such as gas and liquid chromatography−mass spectrometry (GC-MS and LC-MS), are used in untargeted metabolomics. Typically, LC-MS assays yield thousands of features that each represent 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. A significant challenge 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 in reference libraries that were populated 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 nonbiological 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]

Current Landscape and Use of Reference Libraries for Compound Identification. 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 molecular formulas 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, molecular formulas can be inferred using a variety of different software tools, especially if MS/MS and isotope ratio information is included.[25−29]

Many open-access reference libraries exist in the form of compound collections that contain mass, formula and structure information for millions of known or predicted compounds (Table 1). These include PubChem[30] which has nearly 119 million compounds, ChEMBL[31] with 2.1 million compounds, and the US-EPA CompTox Chemicals Dashboard[32] with 1.2 million compounds. All of these support mass and formula searching. However, they also include a large fraction of anthropogenic molecules, making these libraries somewhat more suited for exposomics[33] or environmental studies and less suitable for 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,[34] Lipid Maps[35] lists 45,684 compounds, KEGG[36] denotes 18,784 compounds, and

Table 1. Representative Compound Collection Reference Libraries[a]

library	number of compounds	URL	citation
ChemSpider	>129,000,000	http://www.chemspider.com/	38
PubChem	>119,000,000	https://pubchem.ncbi.nlm.nih.gov/	39
CompTox Chemicals Dashboard	>1,200,000	https://comptox.epa.gov/dashboard/	32
RaMP-DB 2.0	>256,000	https://rampdb.nih.gov/	40
Human Metabolome Database (HMDB)	>248,000	https://hmdb.ca/	34
Metabolomics Workbench	>164,000	https://www.metabolomicsworkbench.org/	41
Chemical Entities of Biological Interest (ChEBI)	>160,000	https://www.ebi.ac.uk/chebi/	42
LipidMaps	>45,000	https://www.lipidmaps.org/databases/lmsd/overview	35
Natural Products Atlas	>33,000	https://www.npatlas.org/	43
MetaboLights	>27,000	https://www.ebi.ac.uk/metabolights/index	44
Kyoto Encyclopedia of Genes and Genomes (KEGG)	>19,000	https://www.genome.jp/kegg/	36
MetaCyc	>16,000	https://metacyc.org/	37

[a]These representative reference libraries function primarily as collections of compounds and include chemical structures, molecular formulae, masses and physicochemical properties, among other data.

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11

Table 2. Representative Reference Libraries of Observable Data[a]

library	number of compounds	number of experimental reference values	URL	citation
Metlin	>860,000	5 million spectra	https://metlin.scripps.edu/	64
NIST2023 EI-MS library	>347,000	>394,000 spectra	https://www.nist.gov/programs-projects/nist23-updates-nist-tandem-and-electron-ionization-spectral-libraries	N/A
NIST2023 RI library	>180,000	>491,000 retention indices	https://www.nist.gov/programs-projects/nist23-updates-nist-tandem-and-electron-ionization-spectral-libraries	N/A
NIST2023 Tandem MS library	>51,000	>2.4 million spectra	https://www.nist.gov/programs-projects/nist23-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/	55
NMRShiftDB	>40,000	>68,000 spectra	https://nmrshiftdb.nmr.uni-koeln.de/	56
Natural Product Magnetic Resonance Database (NP-MRD)	>87,000	>1500 spectra	https://np-mrd.org/	57
FiehnLib RI library	>1200	>1200 retention indices	https://fiehnlab.ucdavis.edu/projects/fiehnlib	59
AllCCS	>2100	>3500 CCS	http://allccs.zhulab.cn/	65
Unified Collision Cross Section Compendium	>1700	>3700 CCS	https://mcleanresearchgroup.shinyapps.io/CCS-Compendium/	60

[a]These representative reference libraries contain listings of compounds and their observable data, such as mass spectra, retention indices, NMR spectra. and CCS values.

MetaCyc[37] includes 16,861 compounds. Many of these databases continue to expand in coverage and content and such databases are much more suitable for traditional metabolomics studies.

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.[45,46] Additional "observable" information is needed to add specificity and increase confidence in tentative compound identifications.[47,48] Generally, the most accessible and reproducible experimental measurements, beyond mass, are spectral 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 ion mobility spectrometry (IMS) data. More recent technology developments allow for the collection of infrared spectra in-line with IMS and MS measurements.[49] 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, MS/MS spectral match alone is insufficient for providing unambiguous matching of metabolites when using 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.[50−52] 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.[53] 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. Representative reference libraries that contain mass spectral data are MassBank.eu, MassBank of North America (MassBank.us), the NIST spectral library,[51] METLIN,[22] and mzCloud. 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.[54] Some of the more popular NMR spectral libraries are the BioMagResBank,[55] NMRShiftDB,[56] NP-MRD,[57] and COLMAR,[58] as well as commercial libraries produced by Bruker and Chenomx. Representative reference libraries that contain RI and/or CCS include: the NIST RI library, the FiehnLib RI library,[59] the Unified CCS Compendium,[60] the Sumner CCS library[61] 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).[62] 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.[63] These data, the largest of their kind, were collected using a single standard chromatographic protocol but have not been validated yet by

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11

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.

Recent Advances in *In Silico* Tools for Expanding Reference Libraries. As can be seen from Table 2, measured observable data are 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.[66] 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.[67] 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 ref 68) 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 example, for ESI-MS/MS spectral prediction, several machine learning methods including MetFrag,[69] CFM-ID,[70] MS-FINDER,[71] ChemDistiller,[72] and MAGMa[73] have appeared. CFM-ID, MS-FINDER and MAGMa in particular have shown excellent performance in terms of spectral prediction accuracy in multiple independent tests.[74,75] For EI-MS spectra, two machine learning methods (CFM-ID-EI[76] and NEIMS[77]) have been described and both perform well. Separately, a quantum mechanical method called QCEIMS[78] has been developed to predict EI-MS spectra and more recently ESI-MS/MS spectra with QCxMS.[79] QCEIMS and QCxMS are significantly slower than the ML methods, but they provide useful insights into the EI and ESI fragmentation processes. We describe advances in predicting NMR spectra, CCS and RT values, and novel metabolite structures (e.g., through biotransformation predictions) in the Supporting Information.

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,[80,81] a critical question is the level of confidence that one has in the identifications proposed. Features detected in untargeted metabolomics analyses are typically identified based on the extent to which their experimental data matches to reference data. In addition to accurate mass (monoisotopic m/z) data and MS/MS spectra obtained from a MS measurement, complementary data from additional analytical measurements (e.g., RT, 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.[47] However, currently there are very few methods for quantifying the ambiguity in a metabolite identification in context of the chemical space being considered, and particularly when extending beyond just MS/MS spectral match. Accurately estimating total FDR in compound identifications is still in its infancy in metabolomics.[52,82,83]

In 2005, a Metabolomics Standards Workshop[84] was convened by the U.S. National Institutes of Health and the Metabolomics Society with the goal of establishing a Metabolomics Standards Initiative (MSI)[85] 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.[47] These MSI-levels have been revised to include additional considerations,[48] or other data types,[86] and focus on specific classes of molecules,[87] but have remained largely unchanged. In 2014, Sumner et al.[88] and Creek et al.[89] 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 identification confidence levels toward 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.[90]

Reporting qualitative 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 Identifiers (PubChem CIDs), Chemical Entities of

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11

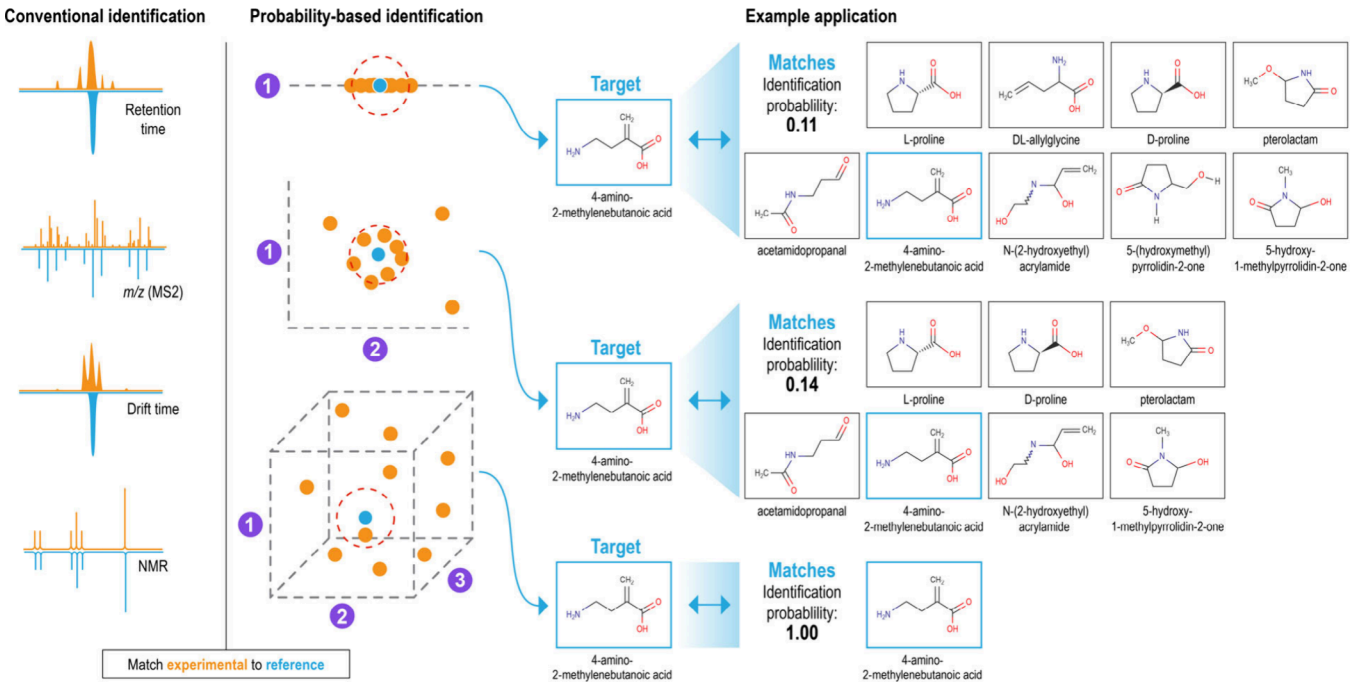


Figure 1. Demonstration of the metabolite identification probability concept using 4-amino-2-methylenebutanoic acid as the target molecule. Conventional metabolite identification (left panel) is based on manual or semiautomated comparison of experimental data to similar data contained in reference libraries, 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 automatable. 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 nonlipid 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 ±5 ppm and ±1% CCS matching (2). In the bottom row, identification is based on the combination of ±5 ppm, ± 1% CCS, and ≥850 cosine similarity score match (3).

Biological Interest (ChEBI), or International Chemical Identifiers (InChIs).[91] 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.[92] 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.[52] The current best alternative to confidence levels is to provide both raw and processed data in public repositories such as the Metabolomics Workbench[41] and MetaboLights[44] 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[47,48,86] and quantitative scoring systems,[88,89] the methods are not easily 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 concept transferable between analytical platforms.

INTRODUCTION TO METABOLITE IDENTIFICATION PROBABILITY

In this section, we introduce the concept of metabolite identification probability and evaluate its implementation using a randomly chosen subset of the Human Metabolome Database (HMDB).[34] The subset was classified into a chemical ontology using the ClassyFire tool,[93] and 27,359 compounds with an invalid chemical classification value ("NA") were excluded, resulting in a nonlipids subset and a lipids subset consisting of 22,077 and 44,537 molecules, respectively (Supplemental Table S1). For each molecule, the protonated mass was calculated from the protonated molecular formula, CCS values were predicted using CCSbase,[94] RTs were predicted using Retip[62] under hydrophilic interaction liquid chromatography (HILIC) conditions for nonlipid molecules or reversed-phase chromatography conditions for lipid molecules, and MS/MS spectra were predicted using CFM-ID 4.0[95] at a "medium" collision energy level of 20 eV. The nonlipid and lipid molecule subsets were placed in separate matrices, together with their protonated mass (m/z), RT, CCS, and MS/MS spectra for each molecule.

Logic Supporting the Concept. Metabolite identification probability represents a first step in moving away from assigning levels of identification confidence or IP-based methods toward a universal, automated method. Importantly, while methods for estimating FDRs for nonpeptide small molecules have been

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11

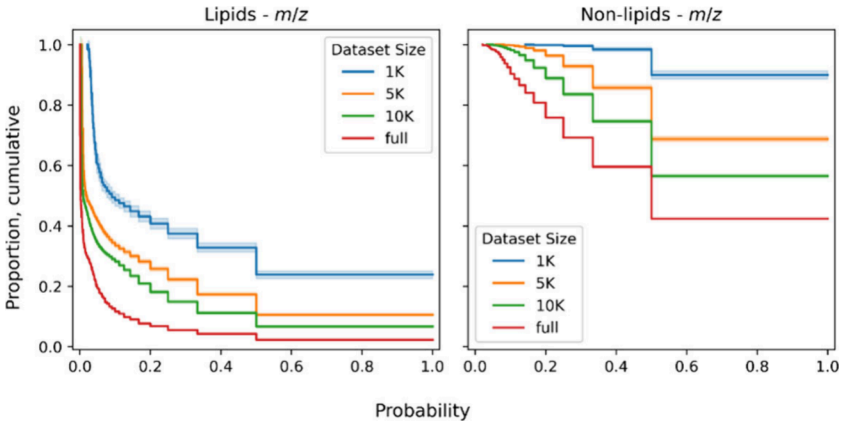


Figure 2. Impact of reference library size on metabolite identification probability. Monte Carlo simulations were performed to randomly draw subsets of the full lipids (left panel) and nonlipids (right panel) reference libraries of size 1K, 5K, and 10K. Match probability is shown on the *x*-axis, and the proportion of compounds in each data set matched within ±5 ppm and with a given probability is shown on the *y*-axis. For example, for nonlipids, a little over 40% of compounds are matched with an identification probability of 100% when matching the full library to itself with a mass tolerance of ±5 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 data set, for which random subsets were not drawn).

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-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)

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 us 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 formulas that correspond to detected isotope patterns in the mass spectrum. The determined molecular formulas can then be searched against an appropriate reference library consisting of known molecular formulas; in our example, we consider a subset of the Human Metabolome Database (HMDB)[34] consisting of 22,077 nonlipid 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 Da, 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 HMDB library subset, we find that there are 9 compounds with the same formula; the probability of the experimentally measured formula $C_5H_9NO_2$ actually being 4-amino-2-methylenebutanoic acid (or any of the 9 candidates) is thus 1/9 or 11% (Figure 1). Now, let us consider a multidimensional 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: ± 5 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 ±5 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 ±5 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 analysis is that the likelihood in complete overlap among property sets for library entries can decrease as dimensionality of the analysis is increased.

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 nonlipid molecules from HMDB, as well as the complementary set of 44,537 lipid molecules (Supplemental Table S2), from the same source, by matching each of the two molecule sets and their calculated/predicted properties to themselves.

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 nonlipids libraries. We evaluated 100 randomly

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11

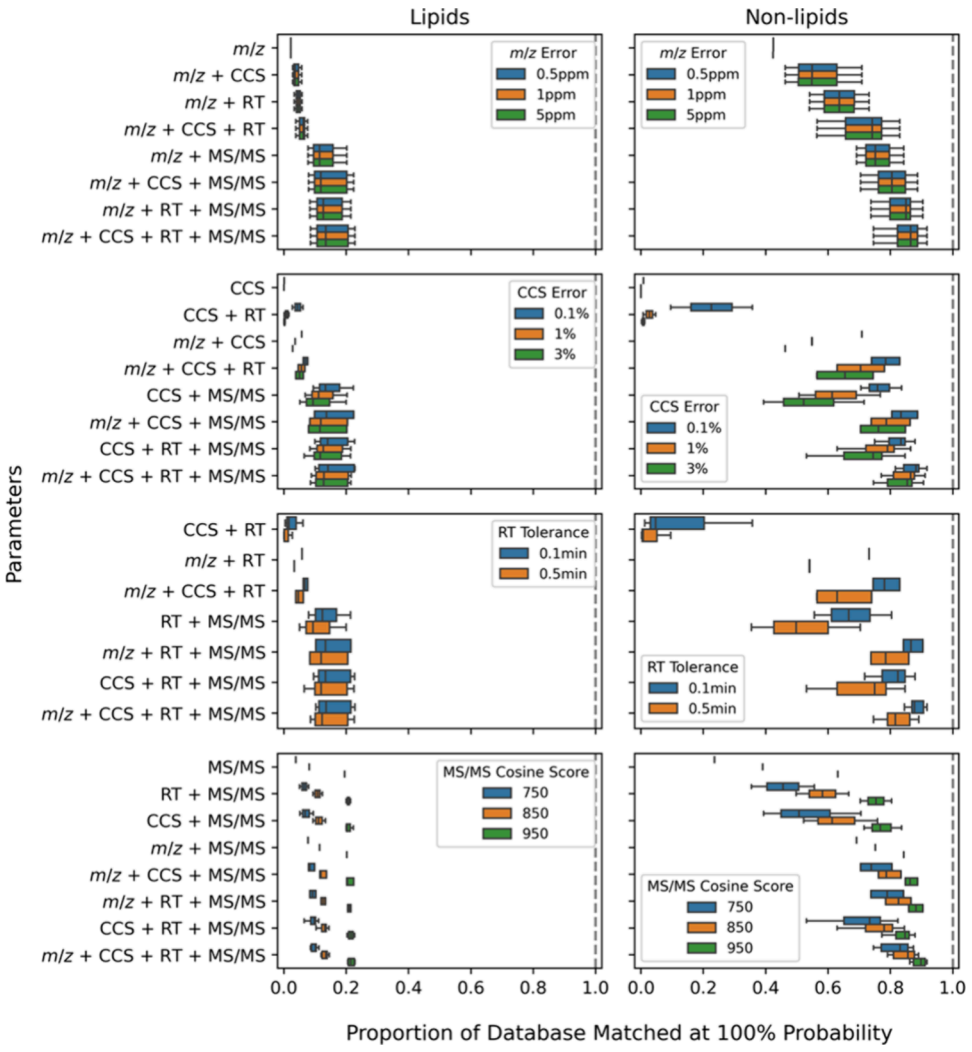


Figure 3. Impact of property match tolerances and dimensionality of experimental analysis on metabolite identification probability for lipids (left) and nonlipids (right). Each boxplot summarizes the fraction of each database that is matched with 100% probability (k = 1) when varying the search tolerances 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 $m/z \pm 0.1$ ppm, $\pm$ 1 ppm, and $\pm$5 ppm; CCS $\pm$ 0.1%, $\pm$ 1%, and $\pm$3%; RT $\pm$ 0.1 min, and $\pm$0.5 min; and MS/MS cosine score $\geq$750, $\geq$ 850, and $\geq$950.

drawn subsets for each library size, matched each subset to itself by mass ($\pm$5 ppm), aggregated results, and compared the number of matches returned per database search (Figure 2). Our results demonstrate that as the size of 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 nonlipids, the impact of reference library size on identification probability is more pronounced for libraries with more heterogeneous content.

Impact of Property Match Tolerances and Dimensionality of Experimental Analysis. We next evaluated the impacts of individual property match tolerances and the dimensionality of the experimental analysis on metabolite identification probability, selecting a range of property match tolerances as might be encountered with a variety of instrumentation (e.g., FTICR vs time-of-flight for mass accuracy). Overall, varying property match tolerance has different impacts on the number of matches with 100% probability depending on the property considered. For instance, the evaluated m/z match thresholds gave rise to little, if any, change in the proportion of matches with 100% probability from both the lipids and nonlipids data sets, either alone or in combination with other properties (Figure 3). We hypothesize 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 $\pm$0.1, 1, and 5 ppm correspond to $\pm$0.00008 Da, $\pm$ 0.0008 Da, and $\pm$0.004 Da, respectively. The resolutions may not differ sufficiently to change the number of matches within each corresponding tolerance significantly.

In contrast to m/z, CCS search tolerance has a more pronounced impact on the number of matches with 100% probability. While searching by CCS alone produces zero or near-zero matches with 100% probability across both lipids and nonlipids data sets, when used in combination with other analytical dimensions, the effect of CCS search tolerance

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11

becomes much more pronounced. In some cases, we observe a 2-fold or even greater increase in the fraction of the data set which can be definitively matched, particularly in the case of lipids (Figure 3). Our simulation 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).[96] Inclusion of CCS in compound matching at this tighter threshold produced marked improvements in the proportion of matches with 100% probability.

Neither of the two RT thresholds evaluated (±0.1 min and ±0.5 min) produced any matches in the database with 100% probability using RT alone for the lipids or nonlipids data sets (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 to 0.1 min correspondingly increases the proportion of matches with 100% probability. While the observed effect is smaller than the impact of CCS, the inclusion of RT still substantially improves matches with 100% probability, 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,[52] 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 nonlipids, MS/MS score alone is the best-performing singular measurement in terms of matching compounds with 100% probability (Figure 3). In contrast to m/z, however, increasing the MS/MS cosine score threshold resulted in significant increases to the proportion of compounds matched with 100% probability in both the lipids and nonlipids libraries. In fact, when matching by MS/MS cosine score alone, 63% of nonlipids 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 compounds matched with 100% probability; in fact, all the best-performing multidimensional search parameter sets include MS/MS.

While the example data and 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,[97] Bayesil[98] and Chenomx,[99] 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.

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 identification error rate and seemingly accurate, and thus overly confident, identification probabilities. One that is too large can result in increased identification error rate due to the addition of compounds that are highly unlikely to be found in such a sample and reduced identification probabilities.[82] 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. More detailed guidelines for appropriate reference library size and composition are discussed in the Supporting Information.

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 nontarget 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.[33,100]

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

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11

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[54] and enabled through workflows such as FragHub.[101] As inspiration for how such sharing might be implemented, the metabolomics community can look to the Universal Protein Knowledgebase (UniProtKB)[102] 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.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.4c04060.

Text describing advances in predicting NMR spectra, CCS and RT values, and novel metabolite structures, text providing detailed guidelines for appropriate reference library size and composition (PDF)

Distribution of ClassyFire superclass annotations of the HMDB subset (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Thomas O. Metz − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0001-6049-3968; Phone: 509-371-6581; Email: thomas.metz@pnnl.gov

Authors

Christine H. Chang − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Vasuk Gautam − *Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada*

Afia Anjum − *Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada*

Siyang Tian − *Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada;* orcid.org/0000-0002-7298-2520

Fei Wang − *Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada; Alberta Machine Intelligence Institute, Edmonton, Alberta T5J 1S5, Canada;* orcid.org/0000-0002-0191-9719

Sean M. Colby − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0002-3193-8267

Jamie R. Nunez − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Madison R. Blumer − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Arthur S. Edison − *Department of Biochemistry & Molecular Biology, Complex Carbohydrate Research Center and Institute of Bioinformatics, University of Georgia, Athens, Georgia 30602, United States;* orcid.org/0000-0002-5686-2350

Oliver Fiehn − *West Coast Metabolomics Center, University of California Davis, Davis, California 95616, United States;* orcid.org/0000-0002-6261-8928

Dean P. Jones − *Clinical Biomarkers Laboratory, Department of Medicine, Emory University, Atlanta, Georgia 30322, United States*

Shuzhao Li − *The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, United States;* orcid.org/0000-0002-7386-2539

Edward T. Morgan − *Department of Pharmacology and Chemical Biology, Emory University School of Medicine, Atlanta, Georgia 30322, United States*

Gary J. Patti − *Center for Mass Spectrometry and Metabolic Tracing, Department of Chemistry, Department of Medicine, Washington University, Saint Louis, Missouri 63105, United States;* orcid.org/0000-0002-3748-6193

Dylan H. Ross − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0009-0005-2943-2282

Madelyn R. Shapiro − *Artificial Intelligence & Data Analytics Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0002-2786-7056

Antony J. Williams − *U.S. Environmental Protection Agency, Office of Research & Development, Center for Computational Toxicology & Exposure (CCTE), Research Triangle Park, North Carolina 27711, United States;* orcid.org/0000-0002-2668-4821

David S. Wishart − *Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada;* orcid.org/0000-0002-3207-2434

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.4c04060

Notes

The authors declare no competing financial interest.

■ 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 multiprogram 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. *Science* **2007**, *316* (5824), 550−551.
(2) Nurse, P.; Hayles, J. *Cell* **2011**, *144* (6), 850−854.
(3) Westerhoff, H. V.; Palsson, B. O. *Nat. Biotechnol.* **2004**, *22* (10), 1249−1252.
(4) Giani, A. M.; Gallo, G. R.; Gianfranceschi, L.; Formenti, G. *Comput. Struct Biotechnol J.* **2020**, *18*, 9−19.
(5) Heather, J. M.; Chain, B. *Genomics* **2016**, *107* (1), 1−8.
(6) Song, Y.; Xu, X.; Wang, W.; Tian, T.; Zhu, Z.; Yang, C. *Analyst* **2019**, *144* (10), 3172−3189.
(7) Aebersold, R.; Mann, M. *Nature* **2016**, *537* (7620), 347−355.
(8) Hashimoto, Y.; Greco, T. M.; Cristea, I. M. *Adv. Exp. Med. Biol.* **2019**, *1140*, 143−154.
(9) Nicholson, J. K.; Lindon, J. C. *Nature* **2008**, *455* (7216), 1054−1056.
(10) Fiehn, O. *Plant Mol. Biol.* **2002**, *48* (1−2), 155−171.
(11) Watson, J. D.; Crick, F. H. *Nature* **1953**, *171* (4356), 737−738.
(12) Franklin, R. E.; Gosling, R. G. *Nature* **1953**, *171* (4356), 740−741.
(13) Schmutz, J.; Wheeler, J.; Grimwood, J.; Dickson, M.; Yang, J.; Caoile, C.; Bajorek, E.; Black, S.; Chan, Y. M.; Denys, M.; et al. *Nature* **2004**, *429* (6990), 365−368.
(14) Lou, D. I.; Hussmann, J. A.; McBee, R. M.; Acevedo, A.; Andino, R.; Press, W. H.; Sawyer, S. L. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (49), 19872−19877.
(15) Mewes, H. W.; Amid, C.; Arnold, R.; Frishman, D.; Guldener, U.; Mannhaupt, G.; Munsterkotter, M.; Pagel, P.; Strack, N.; Stumpflen, V.; et al. *Nucleic Acids Res.* **2004**, *32* (Database issue), 41D−44.
(16) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.
(17) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20* (18), 3551−3567.
(18) Ma, K.; Vitek, O.; Nesvizhskii, A. I. *BMC Bioinformatics* **2012**, *13* (S16), S1.
(19) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4* (3), 207−214.
(20) Dunphy, K.; Dowling, P.; Bazou, D.; O'Gorman, P. *Cancers (Basel)* **2021**, *13* (8), 1930.
(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. *Science* **2016**, *351* (6274), 711−714.
(22) Montenegro-Burke, J. R.; Guijas, C.; Siuzdak, G. *Methods Mol. Biol.* **2020**, *2104*, 149−163.
(23) Frainay, C.; Schymanski, E. L.; Neumann, S.; Merlet, B.; Salek, R. M.; Jourdan, F.; Yanes, O. *Metabolites* **2018**, *8* (3), 51.
(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. *Nat. Commun.* **2021**, *12* (1), 5418.
(25) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2007**, *8*, 105.
(26) Ludwig, M.; Fleischauer, M.; Duhrkop, K.; Hoffmann, M. A.; Bocker, S. *Methods Mol. Biol.* **2020**, *2104*, 185−207.
(27) Pluskal, T.; Uehara, T.; Yanagida, M. *Anal. Chem.* **2012**, *84* (10), 4396−4403.
(28) Duhrkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Bocker, S. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (41), 12580−12585.
(29) Draper, J.; Enot, D. P.; Parker, D.; Beckmann, M.; Snowdon, S.; Lin, W.; Zubair, H. *BMC Bioinformatics* **2009**, *10*, 227.
(30) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. *Nucleic Acids Res.* **2021**, *49* (D1), D1388−D1395.
(31) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100−1107.
(32) 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. *J. Cheminform* **2017**, *9* (1), 61.
(33) Vermeulen, R.; Schymanski, E. L.; Barabasi, A. L.; Miller, G. W. *Science* **2020**, *367* (6476), 392−396.
(34) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; et al. *Nucleic Acids Res.* **2022**, *50* (D1), D622−D631.
(35) Sud, M.; Fahy, E.; Cotter, D.; Dennis, E. A.; Subramaniam, S. *J. Chem. Educ.* **2012**, *89* (2), 291−292.
(36) Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28* (1), 27−30.
(37) 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. *Nucleic Acids Res.* **2018**, *46* (D1), D633−D639.
(38) Pence, H. E.; Williams, A. *J. Chem. Educ.* **2010**, *87* (11), 1123−1124.
(39) 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. *Nucleic Acids Res.* **2016**, *44* (D1), D1202−1213.
(40) Braisted, J.; Patt, A.; Tindall, C.; Sheils, T.; Neyra, J.; Spencer, K.; Eicher, T.; Mathe, E. A. *Bioinformatics* **2023**, *39* (1), No. btac726.
(41) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; et al. *Nucleic Acids Res.* **2016**, *44* (D1), D463−470.
(42) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. *Nucleic Acids Res.* **2016**, *44* (D1), D1214−1219.
(43) 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. *ACS Cent Sci.* **2019**, *5* (11), 1824−1833.
(44) Haug, K.; Cochrane, K.; Nainala, V. C.; Williams, M.; Chang, J.; Jayaseelan, K. V.; O'Donovan, C. *Nucleic Acids Res.* **2019**, *48* (D1), D440−D444.
(45) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2006**, *7*, 234.
(46) Witting, M.; Bocker, S. *J. Sep Sci.* **2020**, *43* (9−10), 1746−1754.
(47) 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. *Metabolomics* **2007**, *3* (3), 211−221.
(48) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. *Environ. Sci. Technol.* **2014**, *48* (4), 2097−2098.
(49) Khanal, N.; Masellis, C.; Kamrath, M. Z.; Clemmer, D. E.; Rizzo, T. R. *Analyst* **2018**, *143* (8), 1846−1852.
(50) Huber, F.; Ridder, L.; Verhoeven, S.; Spaaks, J. H.; Diblen, F.; Rogers, S.; van der Hooft, J. J. J. *PLoS Comput. Biol.* **2021**, *17* (2), No. e1008724.
(51) Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (9), 859−866.
(52) Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O. *Nat. Methods* **2021**, *18*, 1524−1531.
(53) 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. *Anal. Chem.* **2017**, *89* (17), 9048−9055.
(54) 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. *Nat. Biotechnol.* **2016**, *34* (8), 828−837.
(55) Romero, P. R.; Kobayashi, N.; Wedell, J. R.; Baskaran, K.; Iwata, T.; Yokochi, M.; Maziuk, D.; Yao, H.; Fujiwara, T.; Kurusu, G.; et al. *Methods Mol. Biol.* **2020**, *2112*, 187−218.
(56) Steinbeck, C.; Kuhn, S. *Phytochemistry* **2004**, *65* (19), 2711−2717.
(57) Wishart, D. S.; Sayeeda, Z.; Budinski, Z.; Guo, A.; Lee, B. L.; Berjanskii, M.; Rout, M.; Peters, H.; Dizon, R.; Mah, R.; et al. *Nucleic Acids Res.* **2022**, *50*, D665−D677.

(58) Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R. *Anal. Chem.* **2016**, *88* (24), 12411−12418.
(59) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. *Anal. Chem.* **2009**, *81* (24), 10038−10048.
(60) Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. *Chem. Sci.* **2019**, *10* (4), 983−993.
(61) Schroeder, M.; Meyer, S. W.; Heyman, H. M.; Barsch, A.; Sumner, L. W. *Metabolites* **2020**, *10* (1), 13.
(62) Bonini, P.; Kind, T.; Tsugawa, H.; Barupal, D. K.; Fiehn, O. *Anal. Chem.* **2020**, *92* (11), 7515−7522.
(63) Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J. R.; Uritboonthai, W.; Aisporna, A. E.; Chen, E.; Benton, H. P.; Siuzdak, G. *Nat. Commun.* **2019**, *10* (1), 5811.
(64) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther Drug Monit* **2005**, *27* (6), 747−751.
(65) Zhou, Z.; Luo, M.; Chen, X.; Yin, Y.; Xiong, X.; Wang, R.; Zhu, Z. J. *Nat. Commun.* **2020**, *11* (1), 4334.
(66) 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. *Nat. Biotechnol.* **2022**, *40* (12), 1774−1779.
(67) Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; et al. *Nat. Methods* **2018**, *15* (1), 53−56.
(68) 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. *Chem. Rev.* **2021**, *121* (10), 5633−5670.
(69) Wolf, S.; Schmidt, S.; Muller-Hannemann, M.; Neumann, S. *BMC Bioinformatics* **2010**, *11*, 148.
(70) Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D. *Nucleic Acids Res.* **2014**, *42* (Web Server issue), W94−W99.
(71) Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M. *Anal. Chem.* **2016**, *88* (16), 7946−7958.
(72) Laponogov, I.; Sadawi, N.; Galea, D.; Mirnezami, R.; Veselkov, K. A. *Bioinformatics* **2018**, *34* (12), 2096−2102.
(73) Ridder, L.; van der Hooft, J. J.; Verhoeven, S. *Mass Spectrom (Tokyo)* **2014**, *3* (Spec Iss2), S0033.
(74) Schymanski, E. L.; Ruttkies, C.; Krauss, M.; Brouard, C.; Kind, T.; Duhrkop, K.; Allen, F.; Vaniya, A.; Verdegem, D.; Bocker, S.; et al. *J. Cheminform* **2017**, *9* (1), 22.
(75) 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. *Anal Bioanal Chem.* **2020**, *412* (6), 1303−1315.
(76) Allen, F.; Pon, A.; Greiner, R.; Wishart, D. *Anal. Chem.* **2016**, *88* (15), 7689−7697.
(77) Ji, H.; Deng, H.; Lu, H.; Zhang, Z. *Anal. Chem.* **2020**, *92* (13), 8649−8653.
(78) Asgeirsson, V.; Bauer, C. A.; Grimme, S. *Chem. Sci.* **2017**, *8* (7), 4879−4895.
(79) Koopman, J.; Grimme, S. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (7), 1735−1751.
(80) Stanstrup, J.; Broeckling, C. D.; Helmus, R.; Hoffmann, N.; Mathe, E.; Naake, T.; Nicolotti, L.; Peters, K.; Rainer, J.; Salek, R. M.; et al. *Metabolites* **2019**, *9* (10), 200.
(81) Mitchell, J. M.; Chi, Y.; Thapa, M.; Pang, Z.; Xia, J.; Li, S. *PLoS Comput. Biol.* **2024**, *20* (6), No. e1011912.
(82) Matsuda, F.; Shinbo, Y.; Oikawa, A.; Hirai, M. Y.; Fiehn, O.; Kanaya, S.; Saito, K. *PLoS One* **2009**, *4* (10), No. e7490.
(83) Scheubert, K.; Hufsky, F.; Petras, D.; Wang, M.; Nothias, L. F.; Duhrkop, K.; Bandeira, N.; Dorrestein, P. C.; Bocker, S. *Nat. Commun.* **2017**, *8* (1), 1494.
(84) Castle, A. L.; Fiehn, O.; Kaddurah-Daouk, R.; Lindon, J. C. *Brief Bioinform* **2006**, *7* (2), 159−165.
(85) Fiehn, O.; Kristal, B.; van Ommen, B.; Sumner, L. W.; Sansone, S. A.; Taylor, C.; Hardy, N.; Kaddurah-Daouk, R. *OMICS* **2006**, *10* (2), 158−163.
(86) Celma, A.; Sancho, J. V.; Schymanski, E. L.; Fabregat-Safont, D.; Ibanez, M.; Goshawk, J.; Barknowitz, G.; Hernandez, F.; Bijlsma, L. *Environ. Sci. Technol.* **2020**, *54* (23), 15120−15131.
(87) Charbonnet, J. A.; McDonough, C. A.; Xiao, F.; Schwichtenberg, T.; Cao, D.; Kaserzon, S.; Thomas, K. V.; Dewapriya, P.; Place, B. J.; Schymanski, E. L.; et al. *Environ. Sci. Technol. Lett.* **2022**, *9* (6), 473−481.
(88) Sumner, L. W.; Lei, Z.; Nikolau, B. J.; Saito, K.; Roessner, U.; Trengove, R. *Metabolomics* **2014**, *10*, 1047−1049.
(89) 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. *Metabolomics* **2014**, *10*, 350−353.
(90) Alygizakis, N.; Lestremau, F.; Gago-Ferrero, P.; Gil-Solsona, R.; Arturi, K.; Hollender, J.; Schymanski, E. L.; Dulio, V.; Slobodnik, J.; Thomaidis, N. S. *Trac-Trend Anal Chem.* **2023**, *159*, No. 116944.
(91) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. *J. Cheminform* **2013**, *5* (1), 7.
(92) 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. *Nat. Commun.* **2021**, *12* (1), 4771.
(93) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. *J. Cheminform* **2016**, *8*, 61.
(94) Ross, D. H.; Cho, J. H.; Xu, L. *Anal. Chem.* **2020**, *92* (6), 4548−4557.
(95) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. *Anal. Chem.* **2021**, *93* (34), 11692−11700.
(96) 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. *Anal. Chem.* **2019**, *91* (18), 11952−11962.
(97) 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. *Magn. Reson. Chem.* **2023**, *61* (12), 681−704.
(98) Ravanbakhsh, S.; Liu, P.; Bjordahl, T. C.; Mandal, R.; Grant, J. R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; Greiner, R.; Wishart, D. S. *PLoS One* **2015**, *10* (5), No. e0124219.
(99) Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. *Anal. Chem.* **2006**, *78* (13), 4430−4442.
(100) Uppal, K.; Walker, D. I.; Liu, K.; Li, S.; Go, Y. M.; Jones, D. P. *Chem. Res. Toxicol.* **2016**, *29* (12), 1956−1975.
(101) Dablanc, A.; Hennechart, S.; Perez, A.; Cabanac, G.; Guitton, Y.; Paulhe, N.; Lyan, B.; Jamin, E. L.; Giacomoni, F.; Marti, G. *Anal. Chem.* **2024**, *96*, 12489−12496.
(102) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H; Arighi, C. N; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A; Ross, K.; Vinayaka, C R; Wang, Q.; Wang, Y.; Yeh, L.-S.; Zhang, J.; Ruch, P.; Teodoro, D. *Nucleic Acids Res.* **2021**, *49* (D1), D480−D489.

https://doi.org/10.1021/acs.analchem.4c04060
Anal. Chem. 2025, 97, 1−11