

Accurate and Efficient Determination of Unknown Metabolites in Metabolomics by NMR-Based Molecular Motif Identification

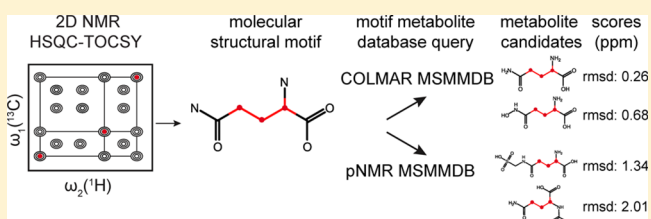
Cheng Wang,[†] Bo Zhang,^{†,∇} István Timári,^{†,○} Árpád Somogyi,[‡] Da-Wei Li,[‡] Haley E. Adcox,[§] John S. Gunn,^{§,#} Lei Bruschiweiler-Li,[‡] and Rafael Brüschweiler^{*,†,‡,||} 

[†]Department of Chemistry and Biochemistry, [‡]Campus Chemical Instrument Center, [§]Department of Microbial Infection and Immunity, and ^{||}Department of Biological Chemistry and Pharmacology, The Ohio State University, Columbus, Ohio 43210, United States

Supporting Information

ABSTRACT: Knowledge of the chemical identity of metabolite molecules is critical for the understanding of the complex biological systems to which they belong. Since metabolite identities and their concentrations are often directly linked to the phenotype, such information can be used to map biochemical pathways and understand their role in health and disease. A very large number of metabolites however are still unknown; i.e., their spectroscopic signatures

do not match those in existing databases, suggesting unknown molecule identification is both imperative and challenging. Although metabolites are structurally highly diverse, the majority shares a rather limited number of structural motifs, which are defined by sets of ¹H and ¹³C chemical shifts of the same spin system. This allows one to characterize unknown metabolites by a divide-and-conquer strategy that identifies their structural motifs first. Here, we present the structural motif-based approach “SUMMIT Motif” for the de novo identification of unknown molecular structures in complex mixtures, without the need for extensive purification, using NMR in tandem with two newly curated NMR molecular structural motif metabolomics databases (MSMMDBs). For the identification of structural motif(s), first, the ¹H and ¹³C chemical shifts of all the individual spin systems are extracted from 2D and 3D NMR spectra of the complex mixture. Next, the molecular structural motifs are identified by querying these chemical shifts against the new MSMMDBs. One database, COLMAR MSMMDB, was derived from experimental NMR chemical shifts of known metabolites taken from the COLMAR metabolomics database, while the other MSMMDB, pNMR MSMMDB, is based on predicted chemical shifts of metabolites of several existing large metabolomics databases. For molecules consisting of multiple spin systems, spin systems are connected via long-range scalar J-couplings. When this motif-based identification method was applied to the hydrophilic extract of mouse bile fluid, two unknown metabolites could be successfully identified. This approach is both accurate and efficient for the identification of unknown metabolites and hence enables the discovery of new biochemical processes and potential biomarkers.



The chemical complexity of living organisms is reflected in their composition of a large number of different metabolites. The human body alone may contain over 100,000 different metabolites, but the majority still needs to be identified and characterized.¹ Such identification is critical to identify potential biomarkers and study new biochemical pathways for the better understanding of biological processes involved in health and disease. The system-wide study of metabolites and pathways, also in relation to the phenotype, is the subject of the field of metabolomics.^{2–4}

Structure determination of novel organic molecules is a standard task in synthetic organic chemistry and natural product research. Analytical methods, such as infrared (IR) and UV–vis spectroscopy, high-resolution mass spectrometry (HRMS), and ¹H and ¹³C NMR spectroscopy, are routinely used.⁵ However, traditional synthetic or natural product characterization requires that a compound has been purified and isolated. In metabolomics, this is often impractical as it can be hard to efficiently isolate a compound at sufficient

concentration among hundreds of molecular species. With respect to NMR, important methodological advances now allow routine characterization of known metabolites in a wide range of different complex mixtures with little or no purification.^{1,6}

Recently, several approaches have been introduced that aim at the de novo characterization and structure determination of metabolites directly in the complex mixture environment.^{7–13} We recently introduced a protocol for the identification of unknown metabolites in metabolomics samples without the need for purification that combines MS, NMR, and cheminformatics.^{14,15} The approach, named SUMMIT MS/NMR, uses accurate mass information, e.g., from Fourier transform ion cyclotron resonance (FTICR), to determine the elemental composition of the metabolites present in the

Received: August 22, 2019

Accepted: November 13, 2019

Published: November 13, 2019

sample. A large pool of chemical compounds is then generated, which are consistent with the MS-derived molecular formulas. The candidate compounds are then filtered against multi-dimensional NMR data, in particular ^1H and ^{13}C chemical shifts that belong to individual spin systems. All candidate compounds are rank-ordered by comparing their predicted NMR chemical shifts with experimentally determined chemical shifts of the unknown compounds in the mixture.

SUMMIT MS/NMR requires that two key conditions are fulfilled: (i) the unknown compound must be present in the pool of candidate structures and (ii) the accuracy of the chemical shift predictor must be sufficiently high to identify the correct compound among potentially many others. In practice, both conditions pose specific challenges. For condition (i), the presence of the unknown metabolite in the pool of structures can be difficult to meet depending on the unknown and how the pool of structures has been generated. Instead, it can be easier and less ambiguous to first establish more general molecular structural properties of the unknown, such as its structural motif(s), before attempting to characterize its full molecular structure. For condition (ii), empirical chemical shift predictors, such as Modgraph/Mnova,^{16,17} are fast, but their accuracy is limited. The average root-mean-square deviations (RMSD) of empirically predicted chemical shifts for a set of representative metabolites are around 0.292 ppm (^1H) and 2.90 ppm (^{13}C) and can be improved to 0.154 ppm (^1H) and 1.93 ppm (^{13}C) by quantum-chemical calculations with multiple scaling¹⁸ at the cost of largely increased computation time. Due to the currently limited accuracy of the chemical shift prediction, the SUMMIT MS/NMR approach returns a potentially large number of compounds as viable candidates, which makes their final verification in terms of their purchase or chemical synthesis followed by spiking experiments in the complex mixture both time-consuming and expensive.

Here, we present an alternative approach, named SUMMIT Motif, for the *de novo* determination of molecular structures of unknowns by first focusing on the determination of molecular structural motifs without the need for any mass spectrometry data. Such motif information represents a key step toward the determination of the full structure. The approach starts out with the identification of ^1H and ^{13}C NMR spin systems to define an unknown metabolite's backbone or contiguous parts thereof. This step is achieved by querying the experimental chemical shifts of the unknown spin system against those of molecular structural motifs (MSMs) in both an experimental and a synthetic MSM chemical shift database. It is shown that this approach is highly effective for MSM identification provided that the MSM of the unknown compound is in fact present in the molecular structural motif metabolomics databases (MSMMDBs). As shown here, this requirement is much easier fulfilled than the condition (i) of SUMMIT MS/NMR described above. The power of the approach is demonstrated by determining unknowns present in mouse bile fluid.

METHODS

Definition of Spin Systems, Molecular Structural Motifs, and COLMAR MSMMDB Curation. For any given molecular structure, MSMs are defined using spin systems as a starting point. As customarily defined, for each proton spin system consisting of N_{H} protons, each ^1H spin is connected to another ^1H spin through no more than 3 bonds. A basic molecular structural motif is then defined by the ^1H spins

together with up to N_{C} carbons (^{13}C or ^{12}C) where each carbon is directly attached to at least one of the ^1H atoms (Figure 1). This “0th shell molecular motif” can then be

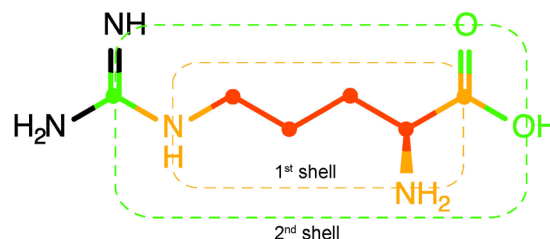


Figure 1. Definition of 1st and 2nd shell molecular motifs based on a spin system. The proton spin system (with N_{H} protons) together with its directly bonded carbon atoms (N_{C} carbons) defines the molecular spin system or “0th shell molecular motif” (red). The “1st shell molecular motif” (dashed orange box) is obtained after inclusion of the heavy atoms that are directly bonded to the spin system (orange). The additional inclusion of heavy atoms that are up to two bonds away from the spin system yields the “2nd shell molecular motif” (dashed green box).

systematically expanded by including additional atoms (N, O, S, P, etc.) that are not directly observable in ^1H and ^{13}C NMR experiments. When including additional heavy atoms that are exactly one bond away, one obtains the “1st shell molecular structural motif”, and when heavy atoms are included that are up to two bonds away, the “2nd shell molecular structural motif” is obtained (Figure 1). The shell order is analogous to the HOSE code used for the empirical prediction of chemical shifts of small molecules.¹⁹ The higher the shell order, the more chemically distinct is the structural motif and the more unique are its ^1H and ^{13}C chemical shifts; however, there is a lower chance that one or several molecules with the same structural motif already exist in a current metabolomics NMR database.

In order to recognize molecular structural motifs that are part of an unknown metabolite, a 1st and 2nd shell molecular structural motif database was generated from the COLMAR small molecule database (experimentally measured in aqueous solution) along with ^1H and ^{13}C chemical shifts of each motif that were assigned to specific groups within each motif. It is called COLMAR MSM Metabolomics Database or COLMAR MSMMDB. COLMAR MSMMDB stores the molecular structures of 1st and 2nd shell MSMs, the parent metabolites of a MSM along with their chemical shifts, their averages, and standard deviations. When only motifs with $N_{\text{C}} > 1$ spin systems were considered, 623 metabolites in the parent COLMAR metabolomics database share 180 unique 1st shell and 397 unique 2nd shell molecular structural motifs (Table S1). Examples of 2nd shell molecular structural motifs are depicted in Figure 2, along with ^1H and ^{13}C chemical shift standard deviations of individual atoms.

The standard deviations of chemical shifts in 2nd shell MSMs are generally well below the average chemical shift errors of NMR chemical shift prediction programs (see Figure S1). This demonstrates that the chemical shifts of 2nd shell MSMs are in most cases considerably more accurate than computationally predicted chemical shifts, which is the reason why 2nd shell MSMs have a better chance to be successfully identified from experimental chemical shifts of unknown metabolites.

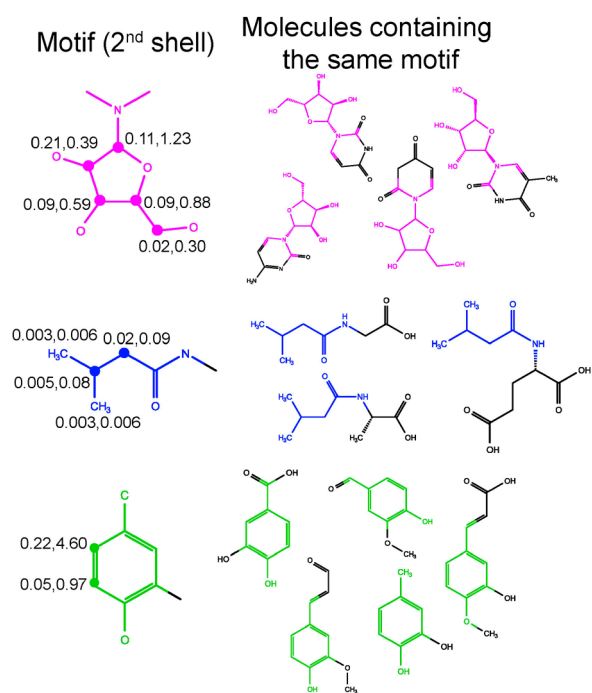


Figure 2. Examples of metabolites with identical molecular structural motifs (in color). The three 2nd shell molecular motifs on the left are highlighted in color. Each row depicts examples of molecules with the same molecular structural motif as the MSM furthest on the left. The experimental NMR chemical shift root-mean-square deviations (RMSDs) at the same C–H positions for all molecules containing the same motif are indicated where the first (second) number is the ^1H (^{13}C) RMSD in units of ppm.

Curation of the Empirically Predicted pNMR MSMMDB. Since the COLMAR MSMMDB features only a subset of MSMs (currently 180 1st shell and 397 2nd shell MSMs) of all possible metabolite MSMs, it is possible that an unknown spin system does not have a good match. For such cases, a MSM database has been built that covers a more diverse set of MSMs from empirically predicted rather than experimental chemical shifts. This database, termed pNMR molecular structural motif metabolomics database (pNMR MSMMDB), consists of MSMs that were extracted from molecules in the HMDB, the Chemical Entities of Biological Interest (ChEBI) database, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.^{20–22} The HMDB is currently the most comprehensive, organism-specific metabolomics database and the largest collection of human metabolites with their chemical structures and biological roles annotated. ChEBI covers both metabolites produced in biological systems and synthetic products that can intervene with living organisms. The KEGG database is one of the most widely used biochemical pathway databases, containing metabolites involved in human diseases and molecular interactions in various organisms. The pNMR MSMMDB, which currently covers 23,697 metabolites with a molecular weight below 800 Da, focuses on motifs of hydrophilic metabolites, which are defined as metabolites with a predicted lipophilicity logP value smaller than 3.0 (as computed by ALOGP,^{23,24} Figure S2). The pNMR MSMMDB contains motifs that overlap with those of COLMAR MSMMDB but with their chemical shifts predicted rather than experimentally determined.

^1H and ^{13}C chemical shifts were computed and stored for each compound in the pNMR MSMMDB using the empirical chemical shift predictor by Modgraph implemented in MestReNova 10.0.1 (Mestrelab Research). The ^1H chemical shift prediction is based on the effects of functional groups that were individually parametrized, whereas the ^{13}C chemical shift prediction is achieved with a HOSE code algorithm. The predicted ^1H and ^{13}C chemical shifts have been sorted into individual spin systems belonging to unique MSMs so that they can be compared directly with experimental ^1H and ^{13}C chemical shifts extracted from experiments. For each MSM, predicted chemical shifts from multiple metabolites are stored separately.

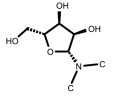

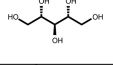
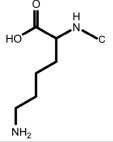
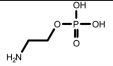
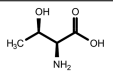
Workflow of MSM-Based Metabolite Identification.

The total workflow of metabolite identification based on molecular structural motifs extracted from NMR spin systems is depicted in Figure S3. After identification of a H–C spin system from 2D/3D TOCSY spectra (see details in the Supporting Information), it is queried against COLMAR MSMMDB. If no hits are returned with RMSD < 2.5 ppm, the spin system information is then queried against pNMR MSMMDB. This two-pronged SUMMIT Motif approach focuses first on the query against the more accurate experimental COLMAR MSMMDB before turning to the larger, but less accurate, pNMR MSMMDB. The MSM hits returned by COLMAR MSMMDB (with RMSD < 2.5 ppm) and the top 15 MSM hits returned by pNMR MSMMDB (with RMSD < 5.0 ppm) are subject to structure determination of unknown metabolites with spiking NMR experiments and/or additional experiments (e.g., 2D HSQMC). Molecular structural motif queries based on COLMAR MSMMDB and pNMR MSMMDB are publicly accessible via the COLMAR suite of web servers (<http://spin.ccic.ohio-state.edu/index.php/motif>).

RESULTS

Evaluation of COLMAR and pNMR MSMMDB in the Identification of Known Molecules in Bile and *E. coli* Extracts. The strategy for the identification of molecular structural motifs (1st or 2nd shell) in unknown metabolites was first tested on known metabolites in bile and *E. coli* cell extracts. There are 26 metabolites in bile and 111 metabolites in *E. coli* cell extract, which could be identified and verified by previously established methods, i.e., 2D HSQC, 2D TOCSY, and 2D HSQC-TOCSY via the COLMAR web server. Each of these known metabolites contains at least one spin system with two or more ^1H spins and their directly attached ^{13}C spins. The ^1H and ^{13}C chemical shifts of each spin system were queried against the COLMAR and pNMR MSMMDB (details regarding spin system matching and scoring are described in the Supporting Information). To be able to treat these spin systems like real unknowns, their true metabolite motifs were intentionally removed from COLMAR MSMMDB. After querying experimental spin systems against COLMAR MSMMDB, both 1st shell and 2nd shell motifs are returned if hits exist within an RMSD cutoff of 5.0 ppm. The motifs with the lowest RMSDs are prioritized for further evaluation. For bile metabolites, the ^1H and ^{13}C chemical shifts of 19 metabolites matched the correct 1st or 2nd shell MSMs from the COLMAR MSMMDB as the top hit. In some cases, a clear distinction between 1st and 2nd shell hits is difficult. For instance, when querying an unknown spin system with the motif of taurine ($\text{SO}_3\text{HCH}_2\text{CH}_2\text{NH}_2$) against the COLMAR

Table 1. Examples of Molecular Structural Motif Identification of Bile Metabolites by COLMAR and pNMR MSMMDB Queries

Chemical shifts of input (^1H , ^{13}C) spin pairs	Motif	Hit that contains true motif returned by COLMAR MSMMDB (Hit rank, type of identified true motif, RMSD (ppm))	Hit that contains true motif returned by pNMR MSMMDB (Hit rank, type of identified true motif, RMSD (ppm))	True Metabolite
(4.120, 86.907) (4.342, 76.337) (4.213, 72.108) (3.802, 63.425) (3.918, 63.385) (5.902, 92.079)		5-Methyluridine (1, 2 nd shell, 0.16)	Beta-D-3-Ribofuranosyluric acid (2, 2 nd shell, 2.08)	Uridine
(7.183, 133.486) (6.888, 118.557)		3-Chlorotyrosine (1, 2 nd shell, 0.20)	N-(1-Deoxy-1-fructosyl)tyrosine (1, 2 nd shell, 1.42)	L-tyrosine
(3.768, 74.784) (3.632, 65.208) (3.610, 73.570)		Adonitol (1, 2 nd shell, 0.79)	Xylitol (1, 2 nd shell, 1.80)	Xylitol
(1.788, 33.111) (2.996, 41.918) (4.135, 57.765) (1.710, 29.096) (1.398, 24.725)		Biocytin (1, 1 st shell, 1.29)	N6-L-homocysteinyln2-L-valyl-L-lysine (3, *2 nd shell, 1.29)	Aspartyl-lysine
(3.966, 62.877) (3.234, 43.195)		Ethanolamine (1, 1 st shell, 1.63)	Phosphoethanolamine (1, 2 nd shell, 1.44)	Phosphoethanolamine
(4.243, 68.598) (3.575, 63.141) (1.318, 22.145)		O-phosphothreonine (1, *2 nd shell, 1.64)	Cyclic N(6)-threonylcarbamoyladenosine (1, 2 nd shell, 1.83)	Threonine

*For these molecules, a partial 2nd shell MSM was identified, since in the true molecule, a hydrogen terminates the 1st shell MSM.

MSMMDB, the identified MSM ($\text{SO}_3\text{HCH}_2\text{CH}_2\text{NH}-\text{CO}-$) is partially a 2nd shell. Overall, the RMSD for the correct hits range between 0.03 and 2.11 ppm (mean value of 0.97 ppm). The RMSD of MSMs of this and other (known) molecules provides a useful benchmark for the range of RMSD values that belong to the correct MSMs. Six metabolites (glycerol, valine, isoleucine, *N*- α -acetyl-L-lysine, leucine, and α -*e*-diaminopimelic acid) did not return any good hits, because no other COLMAR metabolite contains the same 1st or 2nd shell MSMs as these six metabolites. The only spin system whose MSM was misidentified belonged to L-serine where the (incorrect) top hit had a RMSD of 3.21 ppm, confirming that higher RMSDs are generally associated with lower confidence in the returned MSMs. Similarly, after 26 experimental spin systems were queried against the pNMR MSMMDB, the motif of the true metabolites ranks as follows: 1st hits for 12 spin systems, 2nd hits for 4 spin systems, 3rd hits for 4 spin systems, 4th hits for 1 spin system, and 5th to 15th top hits for 5 spin systems. The RMSD of the correct hits range between 0.57 and 2.78 ppm (mean value of 1.78 ppm), which are significantly higher than the RMSD of the hits returned by COLMAR MSMMDB. Examples of MSM identification and representative metabolites with the same MSM returned by COLMAR and pNMR MSMMDB are listed in Table 1. Since pNMR MSMMDB covers a much larger pool of metabolites and MSMs and the chemical shift prediction is less accurate than for COLMAR MSMMDB, the top 15 different MSM hits are considered here as viable candidates when identifying an unknown motif.

For *E. coli* metabolites, the molecular motif of the top hit is 100% correct up to the 1st shell atoms when using an RMSD

threshold of 2.1 ppm. With respect to 2nd shell atoms, 89 out of 111 top hits contain the correct 2nd shell MSMs (true positives) for an RMSD threshold of 5 ppm, whereas 22 top hits contain incorrect 2nd shell molecular motifs (false positives). The quantitative evaluation of true/false positives/negatives for MSM identification is illustrated in the Supporting Information. The number of true and false positive top hits of *E. coli* metabolites with various RMSD thresholds are summarized in Table S2 with the receiver operating characteristic (ROC) curve shown in Figure S4 and an area under the curve (AUC) of 0.851. The number of false positives was reduced when setting the RMSD threshold to 2.1 ppm with 81 true positive 2nd shell MSMs and only 9 false positives. Among the 9 false positives, the true 2nd shell MSMs either ranked as high as 2nd or were not returned at all, because COLMAR MSMMDB did not contain an entry with the same MSM as the true metabolite.

Structure Elucidation of Unknown Metabolites. The determination of MSMs represents a critical step toward the structure elucidation of unknown metabolites, which is demonstrated here for gallbladder bile fluid. We focused on three experimentally extracted spin systems with unknown identity, designated as spin systems A, B, and C. The MSM identification method was applied to identify the unknown molecular motifs belonging to these spin systems followed by identifying the structures of these unknown metabolites. Unknown spin system A has 4 cross-peaks with chemical shifts (δ_{H} , δ_{C}) of (0.883, 20.032), (0.906, 21.543), (2.090, 33.293), and (4.048, 63.569) ppm. After querying against the COLMAR MSMMDB, it best matched the valine-like MSM with structure $\text{CH}_3(\text{CH}_3)\text{CHCH}(\text{COOH})\text{N}-$. When query-

ing against the pNMR MSMMDb, the same valine-like motif that is shared by 32 metabolites was returned. When the top 6 hits with the same molecular motif in the return list were selected and spiking experiments were performed, L-alanyl-L-valine was found to accurately match the unknown spin system (Figure 3). L-Alanyl-L-valine is a dipeptide composed of

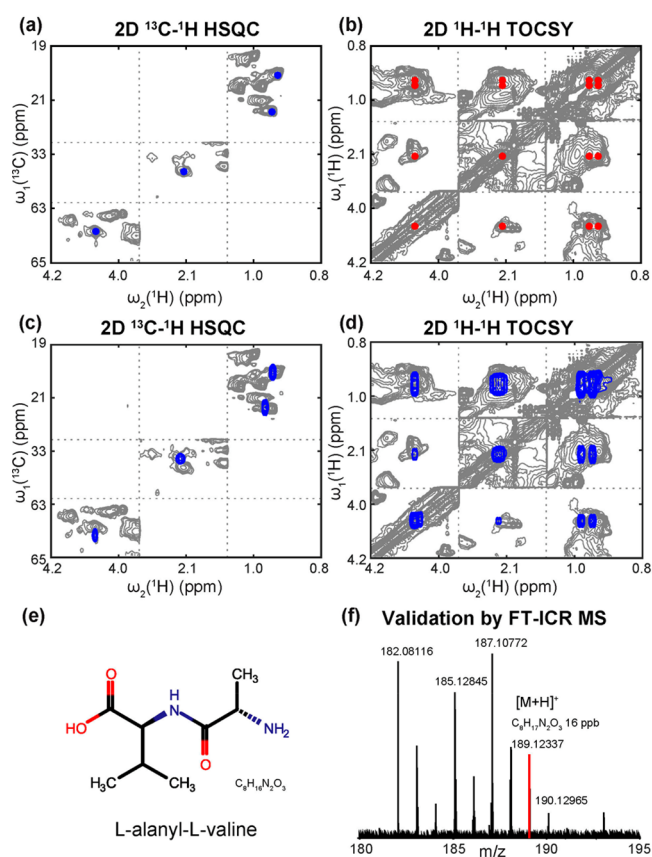


Figure 3. Identification of L-alanyl-L-valine metabolite in mouse bile extracts. Panels a and b: 2D ^{13}C - ^1H HSQC and 2D ^1H - ^1H TOCSY of the unknown spin system A. Panels c and d: overlay of 2D NMR spectra of L-alanyl-L-valine (blue peaks) and bile extracts (gray peaks). The chemical shift agreement confirms the presence of L-alanyl-L-valine in the mouse bile mixture. Panel e: the chemical structure of L-alanyl-L-valine. Panel f: partial FT-ICR mass spectrum of the mouse bile extracts. The red peak with m/z 189.12337 of the $[\text{M} + \text{H}]^+$ adduct is consistent with molecular formula $\text{C}_8\text{H}_{17}\text{N}_2\text{O}_3$ of L-alanyl-L-valine (mass error: 16 ppb).

alanine and valine, which was not previously identified in human tissues or biofluids. Although most dipeptides are relatively short-lived intermediates, some dipeptides are known to have physiological effects, for example, for cell-signaling. The identification of L-alanyl-L-valine in mouse bile provides new information toward the understanding of amino-acid specific pathways for further biological interpretation. The identification and validation of L-alanyl-L-valine is depicted in Figure 3.

Unknown spin system B contains two peaks with chemical shifts (δ_{H} , δ_{C}) of (3.068, 52.507) and (3.558, 37.715) ppm (Figure S5a,b). After querying against the COLMAR MSMMDb, it matched the MSM $-\text{NHCH}_2\text{CH}_2\text{SO}_3\text{H}$, which is also found in taurine. When querying against the pNMR MSMMDb, the same taurine-like MSM was returned that is shared by 96 metabolites. Unknown spin system C

contains 6 peaks with chemical shifts (δ_{H} , δ_{C}) of (0.943 20.558), (1.339 34.342), (1.735 34.345), (1.413 37.477), (2.199 35.428), and (2.296 35.435) ppm (Figure S5c,d). After querying against COLMAR MSMMDb, no hit was returned, indicating that this particular molecular motif does not exist in the COLMAR MSMMDb. By contrast, when querying against pNMR MSMMDb, 33 MSMs were returned that are shared by 49 metabolites. However, the RMSDs of all hits exceeded 2.8 ppm, suggesting that none of the MSM candidates may include the true MSM. The 2D ^{13}C - ^1H HSQC spectrum showed that the two unknown spin systems B and C are part of the same molecule, which were connected via a quaternary carbon peak at 180.820 ppm (Figure S5e). This indicates that the unknown compound has at least two spin systems, and the number of potential unknown compound candidates is lowered from 96 to 25 compounds. The 25 compounds were further separated into four groups based on their second MSMs in addition to MSM ($-\text{NHCH}_2\text{CH}_2\text{SO}_3\text{H}$). When the individual, isolated compounds of the top hit in each MSM group with the lowest RMSD (a total of 4 hits) were selected and purchased and the NMR spiking experiment was performed, taurocholic acid was found to precisely match both unknown spin systems B and C, confirming that they belong to taurocholic acid (Figure S6). The reason that neither COLMAR nor pNMR MSMMDb returned a match for spin system C is that spin system C represents a subspin system of a much larger spin system of taurocholic acid, which contains 19 carbons together with their attached hydrogens. During 120 ms of TOCSY mixing, the magnetization transfer was incomplete and, in addition, a number of cross-peaks overlapped with those of other molecules in the mixture. Therefore, spin system C, which was returned by the maximal clique method, did not correspond to the full spin system, which prevented identification of this motif when querying against pNMR MSMMDb. Although taurocholic acid is a known metabolite of bile, NMR database information for taurocholic acid is available only in DMSO. Because of the dependence of chemical shifts on the solvent (chemical shifts in DMSO are substantially different from those in aqueous condition), taurocholic acid was not part of the COLMAR MSMMDb. These examples illustrate how the molecular structural motif-based method for the identification of unknown metabolites successfully works also for rather large, real-world metabolites.

Coverage of COLMAR Motifs of HMDB. The COLMAR MSMMDb, although established with the motif information from only 632 metabolites, performed remarkably well. The strong performance can be rationalized when comparing the motifs in COLMAR MSMMDb and motifs extracted from the much larger HMDB database. Because a large fraction of the metabolites in the HMDB are hydrophobic metabolites, we focus here on the hydrophilic subset, which includes 13,138 metabolites with a lipophilicity $\log P < 3$ as predicted by ALOGP software.²⁴ All MSMs with 2 or more carbons per spin system were extracted from the hydrophilic HMDB and COLMAR metabolites. The MSM identification results are summarized in Table S1 with 180 COLMAR vs 1924 HMDB 1st shell MSMs and 397 COLMAR vs 4912 HMDB 2nd shell MSMs. The frequency of COLMAR MSMs (nodes and their sizes) is depicted as a network graph in Figure 4, which reflects that MSMs are notably unevenly distributed among metabolites. The most common MSMs are unsaturated carbon-carbon bonds that are part of aromatic ring structures. Many

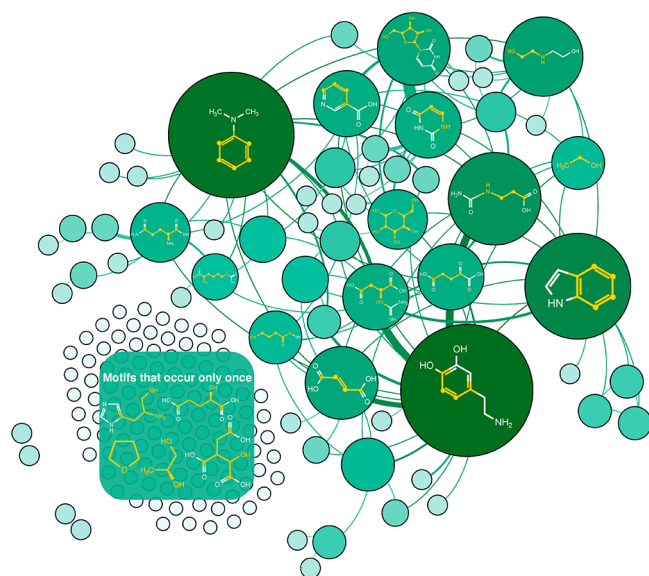


Figure 4. Graphical-theoretical representation of molecular motif clustering of current COLMAR MSMMDB molecules. Each node denotes a molecular motif where the node area is proportional to the number of molecules in the motif. For molecules containing two molecular motifs, their nodes are connected by an edge. The edge thickness (weight) is proportional to the number of molecules that contain both motifs. Representative molecular structures of some of the most abundant molecular motifs are depicted in the graph, and the spin system backbone is highlighted in yellow.

MSMs are frequently found together with other MSMs as parts of larger molecules as is indicated by the many edges of the graph. The top 10 most abundant motifs in the COLMAR and HMDB databases are listed in Table S3 and Table S4. For all 1st shell motifs found in the HMDB database, 37 out of the most frequent 50 motifs are covered by COLMAR (Figure S7). Importantly, the 1st shell COLMAR MSMs cover 10,728 out of 12,506 (85.8%) hydrophilic compounds (with $N_C > 1$ spins for each spin system) of the HMDB, which shows that, despite its much smaller size, COLMAR NMR provides very good coverage of the MSMs of HMDB metabolites. 1778 hydrophilic compounds of the HMDB that contain 1042 different 1st shell MSMs are not covered by COLMAR MSMMDB. However, among the 1042 motifs, only 92 MSMs are present in more than 10 compounds, whereas 92 MSMs represent 5–10 compounds. 858 MSMs are present in fewer than 5 compounds. This motif analysis shows that the vast majority of hydrophilic metabolite motifs not covered by COLMAR MSMMDB are rare motifs.

DISCUSSION

We have established a motif-based method to identify unknown metabolites. The introduction and curation of motif databases are of central importance, whereby the accuracy and precision of the experimental database entries are crucial for the successful identification of MSMs of unknowns. The 1st and 2nd shell atoms of a structural motif sensitively influence the ^1H and ^{13}C chemical shifts and, conversely, experimental ^1H and ^{13}C chemical shifts can be used for the determination of molecular structural motifs. This information offers a path toward the determination of the structure of unknown metabolites. On the basis of the finding presented here that 1st and especially 2nd shell MSMs

generally have chemical shifts that are remarkably well conserved, a MSM NMR database termed COLMAR MSMMDB was established with experimental chemical shifts measured in aqueous solution. The chemical shift accuracy in COLMAR MSMMDB is much better than that of both quantum-chemical¹⁸ and empirical chemical shift prediction used for pNMR MSMMDB. The true MSMs are often found with RMSD < 2.5 ppm, and false MSMs typically have RMSDs > 3.0 ppm. If the RMSD of a MSM is between 2.5 and 3.0 ppm, the application of the pNMR MSMMDB to check whether the same MSM is returned as one of the top 15 hits further increases confidence. Finally, for MSMs (1st shell or 2nd shell) not present in the COLMAR MSMMDB, the more comprehensive but less accurate pNMR MSMMDB consisting of 3512 1st shell MSMs and 7874 2nd shell MSMs with computationally predicted chemical shifts can be used to identify unknown MSMs.

Although the present COLMAR MSMMDB only includes 397 2nd shell MSMs, these MSMs cover a remarkably large number of metabolites. In particular, they represent 10,728 out of 12,506 (85.8%) hydrophilic compounds ($\log P < 3.0$ with $N_C > 1$) of the much larger HMDB, which contains many more metabolites, including many metabolites without experimental NMR chemical shifts and many expected but still unconfirmed metabolites. An interesting question is how many additional MSMs with chemical shifts would need to be added in order to cover all metabolites in the current HMDB. When limiting the MSMs to hydrophilic metabolites only ($\log P < 3$), this can be accomplished with the addition of another 1042 unique MSMs. This is a relatively small number considering that this would cover the MSMs of an additional 1778 HMDB metabolites that are currently not covered by the COLMAR MSMMDB.

The best motifs identified by the COLMAR MSMMDB query can be further developed into complete metabolite candidates. For the examples presented here, a database of potential metabolites was created using a wealth of information from existing databases, such as KEGG, ChEBI, and HMDB. The subset of metabolites that emerge with the correct MSMs and good predicted chemical shift scores represent the candidate molecules that can be purchased (or synthesized) for spiking experiments to confirm their authenticity in the mixture. The SUMMIT Motif approach was successfully illustrated for the identification of two “unknown” metabolites *L*-alanyl-*L*-valine and taurocholic acid in mouse bile fluid. The feasibility of the approach depends on the total concentration of the unknown metabolites, which should exceed ~ 50 – $100 \mu\text{M}$; otherwise, the sample needs to be concentrated first. Because the approach is not designed for high-throughput analysis, it is best applied to a few, selected, representative samples, for example, taken from a large cohort of samples. Also, for unknown compounds that have titratable groups with $\text{pK}_a \sim 7$, even small changes in pH can cause significant changes in chemical shifts that might adversely interfere with motif identification. In principle, NMR pH titration experiments (e.g., via ^{13}C - ^1H HSQC) can help identify such unknowns.

The MSM identification approach presented here is “NMR spin system centric” in the sense that ^1H and ^{13}C spin systems form zeroth shell MSMs, which are then extended to 1st and 2nd shell heavy atoms. As a consequence, the MSMs are only indirectly identifiable by techniques other than NMR. Still, other types of experimental molecular fragment information

can be used, such as metabolite fragments produced by tandem mass spectrometry (MS/MS or MS²), which are routinely used in targeted metabolomics. For the identification of unknown metabolites, MS/MS fragments can be predicted for candidate structures as an additional scoring criterion, thereby further limiting the number of potential metabolites.²⁵ Moreover, the molecular formula of the parent ions, determined from their accurate mass, can serve as an additional filter to narrow down viable candidate compounds that contain the correct MSMs.^{14,15} For instance, if mass information (Figure 3f) was used in addition, L-alanine-L-valine emerged as the only candidate among the six top hits to be subsequently verified by spiking NMR experiments, thereby further speeding up the verification of this unknown metabolite.

CONCLUSIONS

The accurate determination of molecular motifs of unknown metabolites presented here is made possible because of the high quality of NMR chemical shifts of the COLMAR metabolomics database, which was customized primarily using data from the BMRB²⁶ and HMDB.²² Our results suggest that, with the future addition of a modest number of suitably chosen metabolites, the coverage of COLMAR MSMMDB can be substantially broadened to include most real and putative hydrophilic metabolites of the HMDB.

The new COLMAR MSMMDB and pNMR MSMMDB could pave the way for the systematic and efficient determination of motifs and their associated unknown metabolites in a wide range of metabolomics samples. This information will help fill in critical gaps in our understanding of molecular conversions along new metabolomics pathways and their modulations upon internal and external perturbations. The characterization of metabolites with novel MSMs will not immediately benefit from the new COLMAR MSMMDB. It remains a challenge requiring a traditional and, hence, much slower approach for the determination of new natural products relying on extensive purification and comprehensive characterization by the combination of many different analytical techniques. From an NMR perspective, a long-term goal is a universal chemical shift predictor with a similar performance as COLMAR MSMMDB as it would permit the determination of unknowns at a rate that is comparable to the identification of molecular motifs by COLMAR MSMMDB. An empirical predictor with this property would need to be trained on experimental small-molecule chemical shift databases that are much larger than those currently available or require substantially improved quantum-chemical methods for the calculation of chemical shifts. Until then, the systematic addition of high-quality experimental chemical shifts and spin system information of strategically chosen metabolites to NMR metabolomics databases, such as COLMAR MSMMDB, is a practically feasible although time-consuming undertaking.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.9b03849>.

Experimental section, classification of hydrophilic metabolites, spin system identification, matching and scoring, quantitative metric on evaluation of the MSM identification, workflow of SUMMIT Motif, and

examples of top abundant MSMs in COLMAR MSMMDB and HMDB (PDF)

AUTHOR INFORMATION

Corresponding Author

*Tel.: +1-614-644-2083. E-mail: bruschweiler.1@osu.edu.

ORCID

Rafael Brüschweiler: 0000-0003-3649-4543

Present Addresses

[∇]B.Z.: Olaris, Inc., Cambridge, Massachusetts 02139, United States.

[○]I.T.: Department of Inorganic and Analytical Chemistry, University of Debrecen, H-4032 Debrecen, Hungary.

[#]J.S.G.: Center for Microbial Pathogenesis, Abigail Wexner Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, United States.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by a graduate fellowship from Foods for Health (to C.W.), a focus area of the Discovery Themes Initiative at OSU, and the National Institutes of Health (grants R01GM066041 (to R.B.), P30 CA016058, S10 OD018507). All NMR and FT-ICR MS experiments were performed at the Campus Chemical Instrument Center at the Ohio State University.

REFERENCES

- (1) Markley, J. L.; Brüschweiler, R.; Edison, A. S.; Eghbalian, H. R.; Powers, R.; Raftery, D.; Wishart, D. S. *Curr. Opin. Biotechnol.* **2017**, *43*, 34–40.
- (2) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181–9.
- (3) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155–71.
- (4) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–9.
- (5) Pretsch, E.; Bühlmann, P.; Badertscher, M. *Structure determination of organic compounds: tables of spectral data*; Springer: Berlin Heidelberg, 2009.
- (6) Emwas, A. H.; Roy, R.; McKay, R. T.; Tenori, L.; Saccenti, E.; Nagana Gowda, G. A.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; Wishart, D. S. *Metabolites* **2019**, *9*, 123.
- (7) Sanchon-Lopez, B.; Everett, J. R. *J. Proteome Res.* **2016**, *15*, 3405–19.
- (8) Hao, J.; Liebeke, M.; Sommer, U.; Viant, M. R.; Bundy, J. G.; Ebbels, T. M. D. *Anal. Chem.* **2016**, *88*, 2583–2589.
- (9) Chekmeneva, E.; dos Santos Correia, G.; Gomez-Romero, M.; Stamler, J.; Chan, Q.; Elliott, P.; Nicholson, J. K.; Holmes, E. *J. Proteome Res.* **2018**, *17*, 3492–3502.
- (10) Wolfender, J. L.; Nuzillard, J. M.; van der Hoof, J. J. J.; Renault, J. H.; Bertrand, S. *Anal. Chem.* **2019**, *91*, 704–742.
- (11) Djoombou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. *J. Cheminf.* **2019**, *11*, 2.
- (12) Paudel, L.; Nagana Gowda, G. A.; Raftery, D. *Anal. Chem.* **2019**, *91*, 7373–7378.
- (13) Shen, X.; Wang, R.; Xiong, X.; Yin, Y.; Cai, Y.; Ma, Z.; Liu, N.; Zhu, Z. *J. Nat. Commun.* **2019**, *10*, 1516.
- (14) Bingol, K.; Bruschiweiler-Li, L.; Yu, C.; Somogyi, A.; Zhang, F.; Brüschweiler, R. *Anal. Chem.* **2015**, *87*, 3864–70.
- (15) Wang, C.; He, L.; Li, D. W.; Bruschiweiler-Li, L.; Marshall, A. G.; Brüschweiler, R. *J. Proteome Res.* **2017**, *16*, 3774–3786.
- (16) Mestre Lab. *Mnova NMRPredict*; <http://mestrelab.com/software/mnova-nmrpredict-desktop/>.

- (17) Modgraph. *NMRPredict*; http://www.modgraph.co.uk/product_nmr.htm.
- (18) Hoffmann, F.; Li, D. W.; Sebastiani, D.; Brüschweiler, R. *J. Phys. Chem. A* **2017**, *121*, 3071–3078.
- (19) Bremser, W. *Anal. Chim. Acta* **1978**, *103*, 355–365.
- (20) Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28*, 27–30.
- (21) Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcantara, R.; Darsow, M.; Guedj, M.; Ashburner, M. *Nucleic Acids Res.* **2007**, *36*, D344–50.
- (22) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vazquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. *Nucleic Acids Res.* **2018**, *46*, D608–D617.
- (23) Tetko, I. V.; Tanchuk, V. Y. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136–45.
- (24) Thompson, S. J.; Hattotuwigama, C. K.; Holliday, J. D.; Flower, D. R. *Bioinformatics* **2006**, *1*, 237–41.
- (25) Boiteau, R. M.; Hoyt, D. W.; Nicora, C. D.; Kinmonth-Schultz, H. A.; Ward, J. K.; Bingol, K. *Metabolites* **2018**, *8*, 8.
- (26) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Wenger, R. K.; Yao, H. Y.; Markley, J. L. *Nucleic Acids Res.* **2007**, *36*, D402–D408.