

Supporting Information

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

Cheng Wang,¹ Bo Zhang,^{1§} István Timári,¹⁺ Árpád Somogyi,² Da-Wei Li,² Haley E. Adcox,³ John S. Gunn,^{3#} Lei Bruschweiler-Li,² and Rafael Bruschweiler^{1,2,4*}

¹Department of Chemistry and Biochemistry, ²Campus Chemical Instrument Center, ³Department of Microbial Infection and Immunity, ⁴Department of Biological Chemistry and Pharmacology, The Ohio State University, Columbus, Ohio 43210, U.S.A.

Current affiliations:

[§]Olaris, Inc., Cambridge, Massachusetts 02139, U.S.A.

⁺Department of Inorganic and Analytical Chemistry, University of Debrecen, H-4032 Debrecen, Hungary

[#]Center for Microbial Pathogenesis, Abigail Wexner Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, U.S.A.

*To whom correspondence should be addressed:

Rafael Bruschweiler, Ph.D.

CBEC Building, Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio 43210

E-mail: bruschweiler.1@osu.edu

Table of Contents:

Experimental Section

Classification of hydrophilic metabolites based on lipophilicity logP

Single and multiple spin system analysis from 2D and 3D NMR experiments

Spin system matching and scoring

Quantitative metric on the evaluation of the MSM identification

Figure S1. Box plots of chemical shift errors of MSMs.

Figure S2. Histogram of ALOGP values of hydrophilic and hydrophobic metabolites.

Figure S3. Workflow of molecular structural motif (MSM) based unknown metabolite identification.

Figure S4. ROC curve with AUC of 0.851 of true and false positive top hits of *E. coli* metabolites with various RMSD thresholds.

Figure S5. Identification of spin systems of taurocholic acid in mouse bile extracts.

Figure S6. Identification of taurocholic acid in mouse bile extracts.

Figure S7. Distribution of the number of molecules in the 50 most common motifs of hydrophilic compounds of the HMDB.

Table S1. Categorization of COLMAR and HMDB hydrophilic compounds according to their molecular structure motifs.

Table S2. True and false positive top hits of *E. coli* metabolites with various RMSD thresholds (see ROC plot, Figure S4).

Table S3. Molecular structures of top 10 most abundant molecular structural motifs (yellow) of all hydrophilic molecules contained in COLMAR MSMMDB.

Table S4. Molecular structures of top 10 most abundant molecular structural motifs (yellow) of all hydrophilic molecules contained in HMDB.

Experimental Section

Sample preparation

S. Typhimurium was used as a mouse model of *S. Typhi* infection *in vivo*, as currently no animal model for the human-specific *S. Typhi* exists. Naturally resistant NRAMP1(SLC11A1)^{+/+} 129x1/SvJ were fed a lithogenic diet (1% cholesterol and 0.5% cholic acid; Envigo/Harlan Laboratory, IN) for 8 weeks to induce gallstone formation. After completing the diet, mice were intraperitoneally infected with 100 μ L of PBS containing 10⁴ *S. Typhimurium* or 100 μ L of PBS alone. 15 (14) mice were sacrificed 10 days post-infection (PI) for metabolite analysis. About 30 μ L pooled bile was collected from the uninfected mice (PBS) and about 65 μ L pooled bile was collected from the infected mice (*S. Typhimurium*). Both samples were subjected to an immediate metabolite extraction procedure. The freshly collected bile was (sequentially) mixed with 260 μ L ice-cold methanol, 260 μ L ice-cold chloroform, and 195 μ L of ice-cold water and vortexing was applied after each solvent addition. The mixture was then placed on ice for 30 minutes and centrifuged at 5000 x g for 30 min at 4°C for phase separation. The polar phase was then lyophilized and re-dissolved in ice-cold water for subsequent ultra-filtration to remove residual macromolecules. The ultra-filtration step was carried out with an Amicon Ultra 0.5 mL centrifuge filter (MWCO 3 kDa). The filtrate was lyophilized and the powder was reconstituted in 600 μ L D₂O for NMR measurements with 20 mM phosphate buffer to adjust pH to 7-7.4 and 0.1 mM DSS for chemical shift referencing. To prepare the MS sample, 1.5 mg of the dried sample was re-suspended in 200 μ L of H₂O. 10 μ L of the sample was transferred to a new tube followed by 10-fold dilution with 50%/50% (v/v) ACN/H₂O containing 0.1% formic acid. Identification of metabolites was primarily performed on the infected bile samples.

E. coli BL21(DE3) cells were cultured at 37 °C while shaking at 250 rpm in M9 minimum medium with glucose (natural abundance, 5 g/L) added as the sole carbon source. One liter of culture at OD 1 was centrifuged at 5000 x g for 20 min at 4 °C, and the cell pellet was resuspended in 50 mL of 50 mM phosphate buffer at pH 7.0. The cell suspension was then subjected to centrifugation for cell pellet collection. The cell pellet was resuspended in 10 mL of ice-cold water and freeze-thawed three times. The sample was centrifuged at 20,000 x g at 4 °C for 15 min to remove cell debris. Prechilled

methanol and chloroform were sequentially added to the supernatant under vigorous vortexing at an H₂O/methanol/chloroform ratio of 1:1:1 (v/v/v). The mixture was then left at -20 °C overnight for phase separation. Next, it was centrifuged at 4000 x g for 20 min at 4 °C, and the clear upper hydrophilic phase was collected and subjected to rotary evaporation to reduce the methanol content. Finally, the sample was lyophilized. The NMR sample was prepared by dissolving the dry sample in 200 µL of D₂O with 20 mM phosphate buffer to adjust pH to 7-7.4 and 0.1 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) for chemical shift referencing and then transferred to a 3 mm NMR tube.

Experiments and data processing

All NMR spectra were collected on a Bruker Advance III 850 MHz spectrometer equipped with a cryogenically cooled TCI probe at 298K. The 2D ¹³C-¹H HSQC spectra of mice bile samples and *E. coli* cell extracts were collected with 512 x 1024 (N₁ x N₂) complex points along the two dimensions with 32 scans per increment. The spectral widths along the ¹³C and ¹H dimensions were 34206.23 and 10204.08 Hz, respectively, and the transmitter frequency offsets were 75.00 and 4.70 ppm, respectively.

The 2D ¹³C-¹H HSQC-TOCSY spectra were collected with 512 x 2048 (N₁ x N₂) complex points along the two dimensions with 32 scans per increment. The spectral widths along the ¹³C and ¹H dimensions were 34206.23 and 10204.08 Hz, respectively. The transmitter frequency offsets were 75.00 and 4.70 ppm, respectively.

The 2D ¹H-¹H TOCSY spectra were collected with 512 x 2048 (N₁ x N₂) complex points along the two dimensions with 8 scans per increment. The spectral widths along indirect ¹H and direct ¹H dimensions were 10201.97 and 10204.08 Hz, respectively, and the transmitter frequency offset was 4.70 ppm.

The 3D ¹³C-¹H HSQC-TOCSY spectra were collected with 64 x 100 x 1048 (N₁ x N₂ x N₃) complex points along the three dimensions with 8 scans per increment. The spectral widths along the indirect ¹³C, indirect ¹H and direct ¹H dimensions were 34204.76, 10204.09 and 10204.08 Hz, respectively, and the transmitter frequency offsets were 75.00, 4.70 and 4.70 ppm, respectively.

The isotropic mixing times for 2D ^{13}C - ^1H HSQC-TOCSY, 2D ^1H - ^1H TOCSY and 3D ^{13}C - ^1H HSQC-TOCSY were 120, 117 and 117 ms, respectively. The relaxation delay (d1) was 1.5 s.

The 2D ^{13}C - ^1H HSQMBC spectra were collected with 320 x 2048 (N_1 x N_2) complex points along the two dimensions with 64 scans per increment. The spectral widths along the ^{13}C and ^1H dimensions were 40621.32 and 10204.08 Hz, respectively and the transmitter frequency offsets were 90.00 and 4.70 ppm, respectively. The multiple-bond coupling constant was set to 6 Hz. All spectra were zero-filled, Fourier-transformed, and phase- and baseline-corrected using NMRPipe.

FT-ICR MS spectra were collected on a 15 Tesla FT-ICR MS experiment using methods established previously.¹ Briefly, after initial calibration by a standard amino acid mixture, electrospray ionization (ESI) mode was selected and the mass range (m/z) was set to 50 - 3000. After tuning, acquisition and calibration, both positive and negative ion mode mass spectra of the background and metabolite mixture were collected.

The FT-ICR mass spectra were calibrated and analyzed based on common compounds in the metabolite mixture. The mass peak list (m/z) from the mass range 100-1000 was generated with the signal to noise ratio set to 10. The background mass peaks were removed. For each mass peak, all the mass peaks (m/z) were converted to accurate masses with possible adducts.

Classification of hydrophilic metabolites based on lipophilicity logP

Molecular structural motif clustering was applied to hydrophilic metabolites contained in the HMDB. The selection criterion for hydrophilic vs. hydrophobic metabolites was based on the quantitative molecular hydrophobicity (lipophilicity) measure $\log_{10}(\text{P})$ (or $\log\text{P}$) where P is the 1-octanol/water partition coefficient of a given metabolite.² For this purpose, lipophilicities were predicted by ALOGP, which is a widely used computational estimator of $\log\text{P}$, for 933 known metabolites with both hydrophilic and hydrophobic character taken from the COLMAR database, HMDB, and BMRB. 723 compounds whose NMR data were measured in D_2O are pre-classified as hydrophilic compounds and 210 compounds whose NMR data were measured in CDCl_3 are pre-classified as

hydrophobic compounds. Based on the ALOGP distribution (**Figure S2**) an ALOGP criterion < 3.0 was chosen for hydrophilic metabolites.

Single and multiple spin system analysis from 2D and 3D NMR experiments

Entire spin systems of individual molecules are extracted from 2D ^1H - ^1H TOCSY, 2D ^{13}C - ^1H HSQC, and 3D ^{13}C - ^1H HSQC-TOCSY spectra directly applied to the complex mixture of interest. This is automatically achieved by applying methods developed previously in our lab using graph theory and maximal clique analysis.³ Specifically, each directly bonded ^{13}C - ^1H cross-peak in a 2D HSQC is defined as a node of a mathematical graph. Edges between the nodes correspond to connectivities between pairs of ^{13}C - ^1H cross-peaks observed in TOCSY-type spectra (2D TOCSY, 2D HSQC-TOCSY, and 3D HSQC-TOCSY). The graph is then subjected to maximal clique analysis using the Bron-Kerbosch algorithm where individual cliques correspond to separate spin systems.^{1,3}

Considering that molecules can be composed of multiple motifs, once individual motifs have been identified, the connectivity of multiple MSMs can be further explored by additional NMR experiments. 2D heteronuclear multiple bond correlation (HMBC) experiment is widely used to measure long-range heteronuclear coupling constants for the stereochemical and conformational analysis of biologically active natural products,⁴ and it is also a promising tool for the structure elucidation of metabolites and natural products in complex mixtures.^{5,6} The heteronuclear single quantum multiple bond correlation (HSQMBC) experiment⁷ detects long-range heteronuclear correlations through $^nJ(\text{CH})$ ($n>1$) couplings ($\sim 2 - 10$ Hz), which makes the identification of quaternary carbons possible as well. Connectivity information between separate spin systems within a molecule can hence be retrieved via the 2D ^{13}C - ^1H PIP-HSQMBC spectrum.⁸ Therefore, after extracting individual spin systems of each compound from 2D and 3D TOCSY-type spectra, a 2D ^{13}C - ^1H HSQMBC experiment is performed to identify connectivities between different spin systems to establish whether they belong to the same compound.

Spin system matching and scoring

The ^1H , ^{13}C chemical shifts of each compound in COLMAR MSMMDB and pNMR MSMMDB were compared with each experimental spin system with the same number of

spins. The weighted matching algorithm, known as the Hungarian method using the Munkres assignment algorithm, was applied to find the closest matching peak pairs between the experimental and predicted spin systems.⁹ The corresponding chemical shift root-mean-square deviation (RMSD) was calculated between each experimentally determined spin system and each candidate compound according to:

$$RMSD = \left\{ \sum_{i=1}^N [(C_{i,exp} - C_{i,pred})^2 + ((H_{i,exp} - H_{i,pred}) \times 10)^2] / 2N \right\}^{1/2} \quad (1)$$

X_{exp} are the experimental chemical shifts, X_{pred} are the predicted chemical shifts, and N is the number of HSQC cross-peaks of the spin system. A scaling factor of 10 is used to normalize the effects of ^{13}C and ^1H chemical shifts on the overall RMSD by correcting for the different chemical shift ranges of ^{13}C vs. ^1H nuclei. For each experimentally determined spin system, all compounds that fulfill the chemical shift RMSD cutoff < 5 ppm are rank-ordered with the smallest RMSD appearing first.

Quantitative metric on the evaluation of the MSM identification

To quantitatively evaluate the MSM identification result, the true/false positive/negative results are described here. A true positive (TP) is defined as a top hit returned by COLMAR MSMMDB that contains the true metabolite MSM, while a false positive (FP) does not contain the true metabolite MSM. A true negative (TN) results when the true metabolite MSM does not exist in COLMAR MSMMDB and no hit is returned by COLMAR MSMMDB. A false negative (FN) results when no hit is returned by COLMAR MSMMDB, while the true metabolite MSM exists in COLMAR MSMMDB. The true positive rate (TPR) is calculated based on $TP/(TP + FN)$. The false positive rate (FPR) is calculated based on $FP/(FP + TN)$.

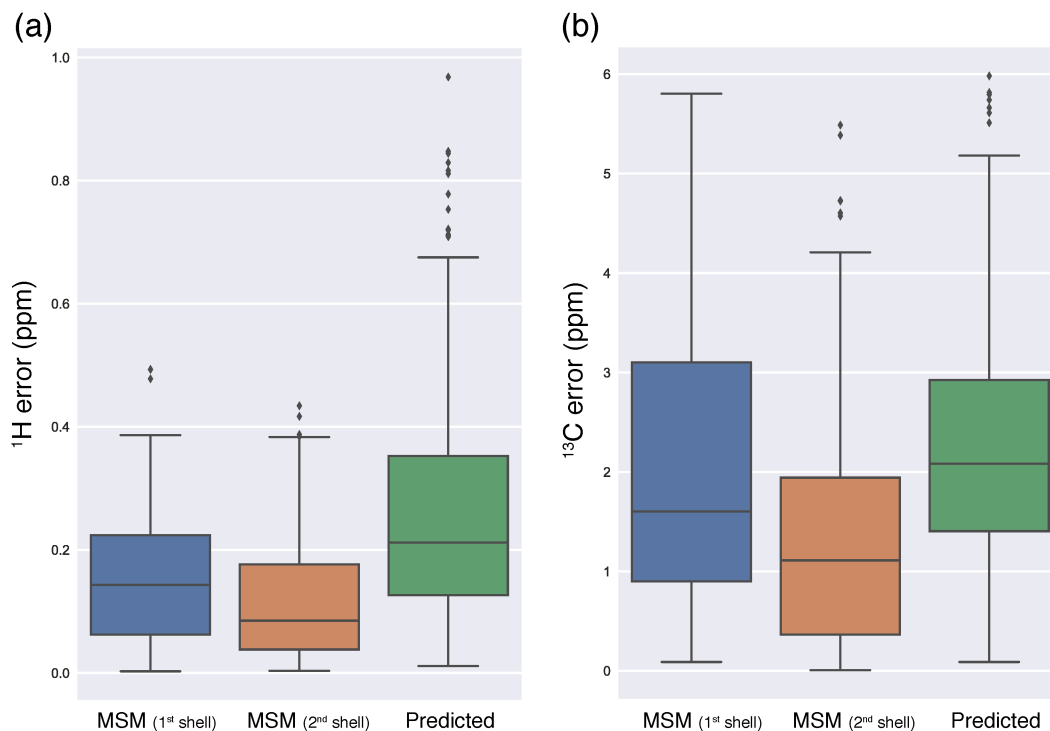


Figure S1. Box plots of chemical shift errors of MSMs. The chemical shift errors of MSMs are calculated based on the average chemical shift of each COLMAR MSM (with occurrence more than 3) versus the true chemical shift of each COLMAR MSM in each compound. The chemical shift errors of prediction are calculated based on the predicted chemical shift of each COLMAR MSM (with occurrence > 3) versus the true chemical shift of each COLMAR MSM in each compound. The statistics of ¹H, ¹³C chemical shift errors (ppm) of 1st shell, 2nd shell MSM and predictions are summarized below. The averages are (0.158, 2.03), (0.115, 1.32) and (0.269, 2.29). The medians are (0.143, 1.60), (0.085, 1.11) and (0.211, 2.08). The first quartiles (Q1) are (0.062, 0.90), (0.038, 0.36) and (0.126, 1.40). The third quartiles (Q3) are (0.224, 3.10), (0.176, 1.94) and (0.352, 2.92). The interquartile ranges (IQR, Q3 - Q1) are (0.162, 2.20), (0.138, 1.58) and (0.226, 1.52). The “minimum” (lower bound, Q1 - 1.5 x IQR) are (0.003, 0.089), (0.003, 0.006) and (0.011, 0.089). The “maximum” (upper bound, Q3 + 1.5 x IQR) are (0.47, 5.80), (0.384, 4.31) and (0.692, 5.20). We only consider common MSM with occurrence > 3 and spin system size ≥ 2 in COLMAR, which Modgraph tends to predict well. Single CH,

CH₂ and CH₃ groups in a molecule (e.g., in adenosine) were not included for calculating prediction error.

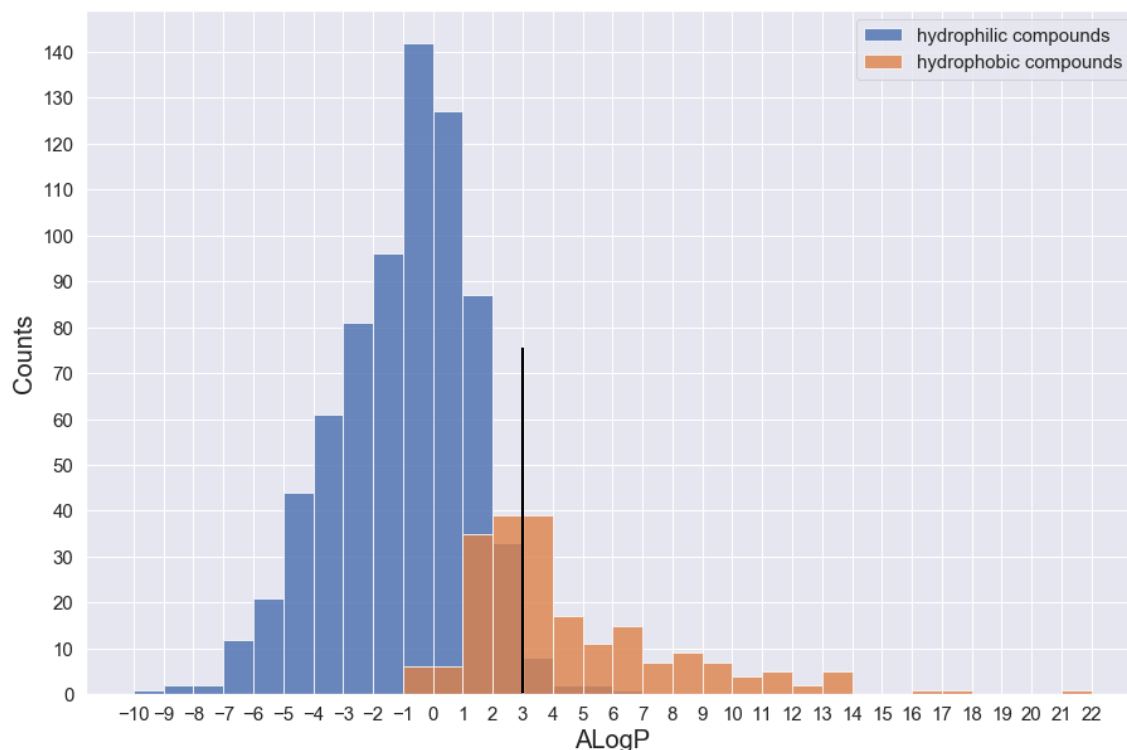


Figure S2. Histogram of ALOGP values of hydrophilic and hydrophobic metabolites. The purpose of this figure is to determine a lipophilicity criterion to accurately classify metabolites without experimental NMR data into hydrophilic and hydrophobic metabolites. This allows one to limit the pool of potential metabolites found in aqueous metabolomics samples. 933 hydrophilic and hydrophobic metabolites from the COLMAR, HMDB and BMRB databases were classified according to their predicted lipophilicity ALOGP values, which are estimates of the experimental $\log_{10}(P)$ values. Metabolites whose NMR spectra were measured in aqueous solution are indicated in blue (“hydrophilic metabolites”) and metabolites measured in organic CDCl_3 solvent are indicated in red color (“hydrophobic metabolites”). When setting the threshold for ALOGP values to 3.0, 710 out of 723 (98.2%) hydrophilic compounds are classified as hydrophilic and 86 out of 210 (41.0%) hydrophobic compounds as hydrophobic. For pNMR MSMMDB only compounds with $\text{ALOGP} < 3.0$ were included from the parent databases (HMDB, KEGG, ChEBI).

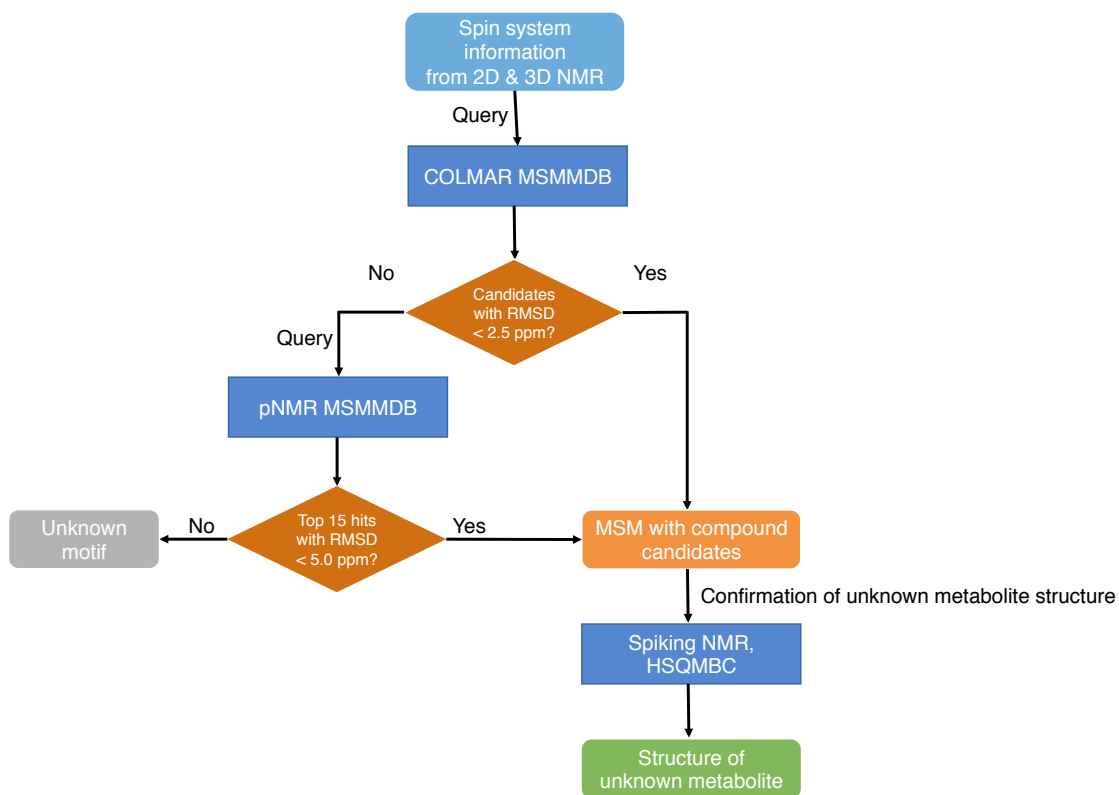


Figure S3. Workflow of molecular structural motif (MSM) based unknown metabolite identification.

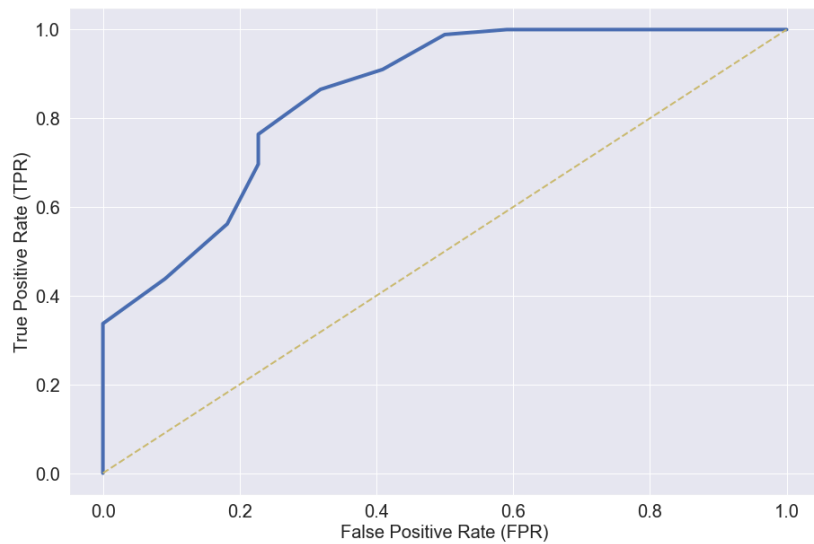


Figure S4. ROC curve with AUC of 0.851 of true and false positive top hits of *E. coli* metabolites with various RMSD thresholds.

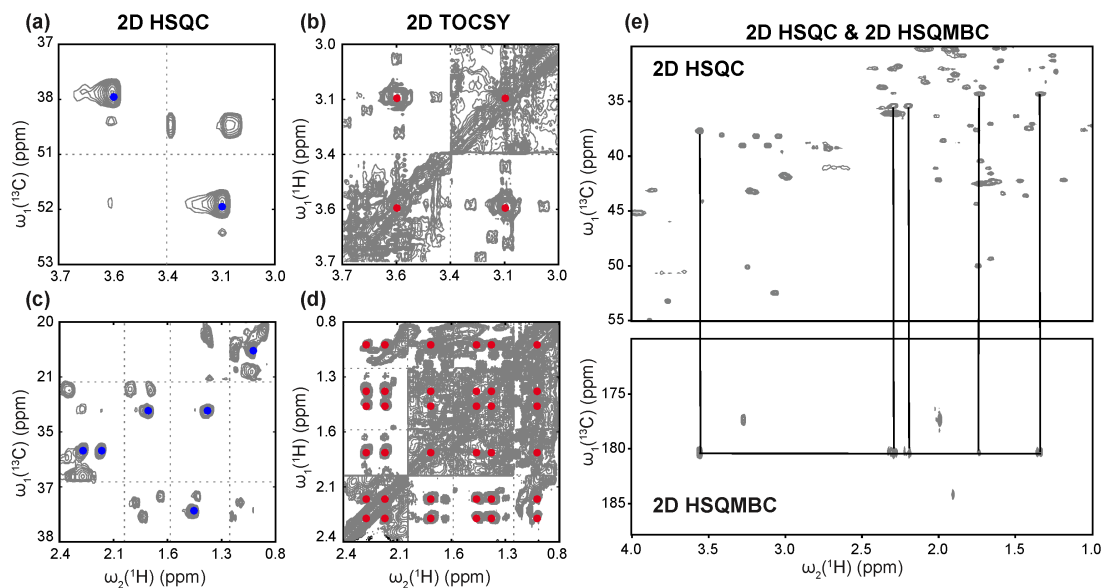


Figure S5. Identification of spin systems of taurocholic acid in mouse bile extracts. Panels a and b: 2D ^{13}C - ^1H HSQC and 2D ^1H - ^1H TOCSY of spin system B. Panels c and d: 2D ^{13}C - ^1H HSQC and 2D ^1H - ^1H TOCSY of spin system C. Panel e: 2D ^{13}C - ^1H HSQMBC indicates spin system B and C are connected via the quaternary carbon that has a signal at 180.82 ppm, suggesting that the two spin systems belong to the same unknown compound.

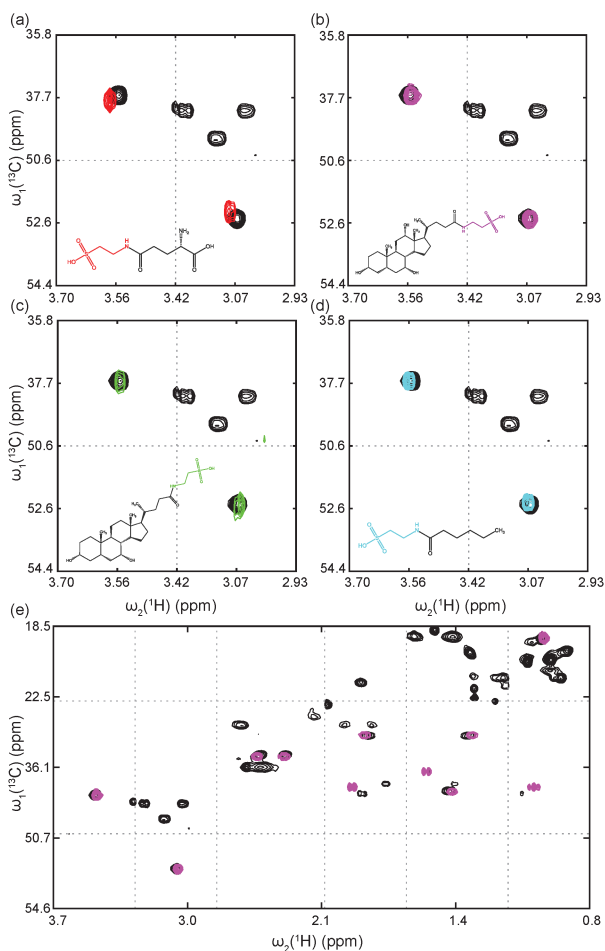


Figure S6. Identification of taurocholic acid in mouse bile extracts. Panels a, b, c, d: overlay of ^{13}C - ^1H HSQC spectra of taurine-like motifs $-\text{NHCH}_2\text{CH}_2\text{SO}_3\text{H}$ present in the top 4 hit molecules (color coded in red, magenta, green and cyan) and spin system B in the bile extracts (2 grey peaks underneath the colored peaks). Since taurocholic acid also matched the spin system C of the unknown metabolite (**Figure S5**), a reference spectrum of taurocholic acid was measured, which matched the HSQC cross-peak well, thereby confirming the presence of taurocholic acid in bile (Panel e). Panel e: overlay of 2D ^{13}C - ^1H HSQC spectra of taurocholic acid (magenta peaks) and the mouse bile extracts (gray peaks).

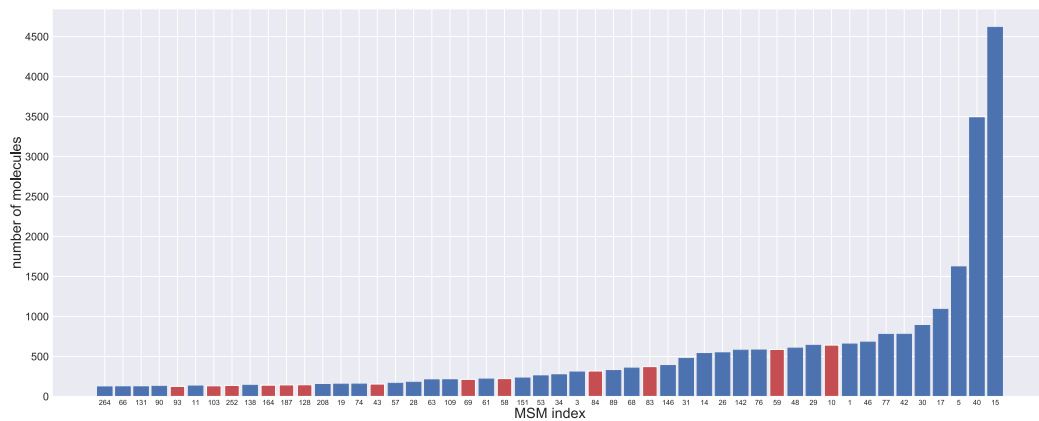


Figure S7. Distribution of the number of molecules in the 50 most common motifs of hydrophilic compounds of the HMDB. A blue bar denotes that the motif exists both in the COLMAR MSMMDB and HMDB of hydrophilic compounds, whereas a red bar denotes that the motif only exists in the HMDB database of hydrophilic compounds.

Table S1. Categorization of COLMAR and HMDB hydrophilic compounds according to their molecular structure motifs.

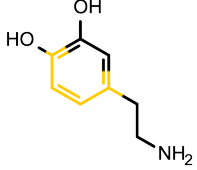
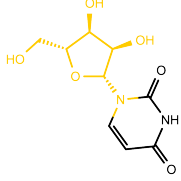
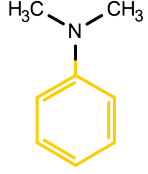
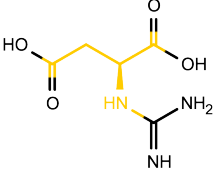
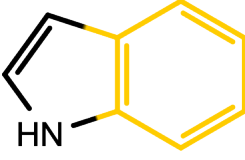
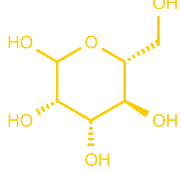
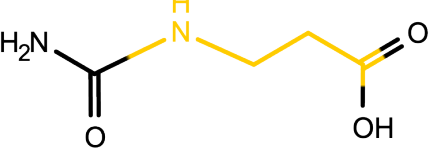
Compound source	COLMAR	HMDB
Total number of compounds	720	13138 ^a
Minimal spin system size	> 1 ¹ H- ¹³ C spin pair per spin system (i.e. N _C > 1)	> 1 ¹ H- ¹³ C spin pair per spin system (i.e. N _C > 1)
Number of compounds categorized	623	12506
Number of 1 st shell motifs	180	1924
Number of 2 nd shell motifs	397	4912

^a The compounds were retrieved from HMDB version 4.0. All metabolites (detected and expected metabolites) with ALOGP < 3.0 were included (see Figure S2); predicted HMDB metabolites were not included.

Table S2. True and false positive top hits of *E. coli* metabolites with various RMSD thresholds (see ROC plot, Figure S4).

RMSD threshold (ppm)	True positive	False positive
0.3	30	0
0.6	39	2
0.9	50	4
1.2	62	5
1.5	68	5
1.8	77	7
2.1	81	9
2.4	88	11
2.7	89	13
3.0	89	16
3.3	89	19
3.6	89	21
3.9	89	22

Table S3. Molecular structures of top 10 most abundant molecular structural motifs (yellow) of all hydrophilic molecules contained in COLMAR MSMMDB.

Number of molecules with the same motif	Molecular structural motif (MSM)
78	
41	
38	
34	
32	
27	
22	

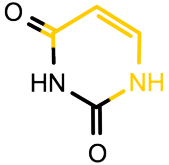
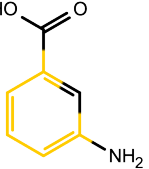
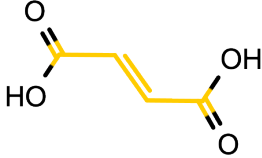
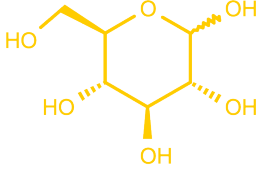
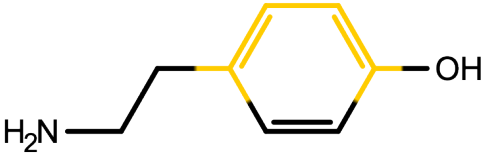
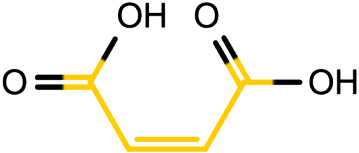
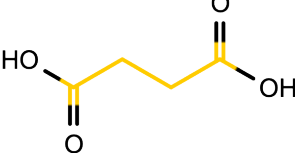
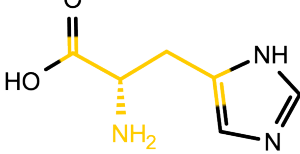
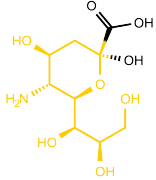
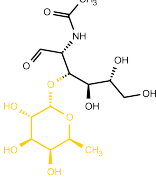
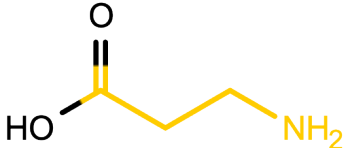
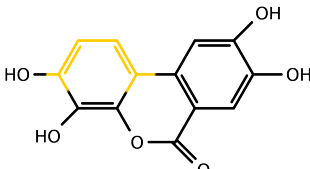
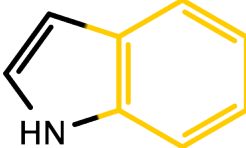
22	 <chem>O=C1NC(=O)NC(=O)N1</chem>
18	 <chem>NC1=CC=C(C(=O)O)C=C1</chem>
18	 <chem>OC(=O)/C=C/C(=O)O</chem>

Table S4. Molecular structures of top 10 most abundant molecular structural motifs (yellow) of all hydrophilic molecules contained in HMDB.

Number of molecules with the same motif	Molecular structural motif (MSM)
4556	
3375	
1600	
1075	
851	
776	
772	

654	 <p>Chemical structure of L-glutamic acid, showing a carboxylic acid group (HO-C=O) and an amino group (NH₂) on a five-carbon chain.</p>
642	 <p>Chemical structure of Gallic acid, a trihydroxybenzoic acid derivative, consisting of a central benzene ring with three hydroxyl groups and a carboxylic acid group.</p>
623	 <p>Chemical structure of Indole, a bicyclic aromatic heterocycle consisting of a benzene ring fused to an imidazole ring.</p>

References

- (1) Wang, C.; He, L.; Li, D. W.; Bruschweiler-Li, L.; Marshall, A. G.; Bruschweiler, R., Accurate Identification of Unknown and Known Metabolic Mixture Components by Combining 3D NMR with Fourier Transform Ion Cyclotron Resonance Tandem Mass Spectrometry. *J. Proteome Res.* **2017**, *16*, 3774-3786.
- (2) Leo, A.; Hansch, C.; Elkins, D., Partition coefficients and their uses. *Chem. Rev.* **1971**, *71*, 525-616.
- (3) Li, D. W.; Wang, C.; Bruschweiler, R., Maximal clique method for the automated analysis of NMR TOCSY spectra of complex mixtures. *J. Biomol. NMR* **2017**, *68*, 195-202.
- (4) Parella, T.; Espinosa, J. F., Long-range proton-carbon coupling constants: NMR methods and applications. *Prog. Nucl. Magn. Reson. Spectrosc.* **2013**, *73*, 17-55.
- (5) Dumas, M. E.; Canlet, C.; André, F.; Vercauteren, J.; Paris, A., Metabonomic assessment of physiological disruptions using ¹H-¹³C HMBC-NMR spectroscopy combined with pattern recognition procedures performed on filtered variables. *Anal. Chem.* **2002**, *74*, 2261-73.
- (6) Bakiri, A.; Hubert, J.; Reynaud, R.; Lambert, C.; Martinez, A.; Renault, J. H.; Nuzillard, J. M., Reconstruction of HMBC Correlation Networks: A Novel NMR-Based Contribution to Metabolite Mixture Analysis. *J. Chem. Inf. Model.* **2018**, *58*, 262-270.
- (7) Williamson, R. T.; Márquez, B. L.; Gerwick, W. H.; Kövér, K. E., One- and two-dimensional gradient-selected HSQMBC NMR experiments for the efficient analysis of long-range heteronuclear coupling constants. *Magn. Reson. Chem.* **2000**, *38*, 265-273.
- (8) Castañar, L.; Saurí, J.; Williamson, R. T.; Virgili, A.; Parella, T., Pure in-phase heteronuclear correlation NMR experiments. *Angew. Chem. Int. Ed. Engl.* **2014**, *53*, 8379-82.
- (9) Kuhn, H. W., The Hungarian method for the assignment problem. *Naval research logistics quarterly* **1955**, *2*, 83-97.