Journal of proteome research

Article

pubs.acs.org/jpr

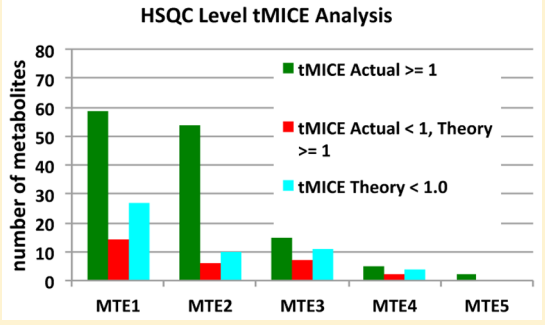
New Methodology for Known Metabolite Identification in Metabonomics/Metabolomics: Topological Metabolite Identification Carbon Efficiency (tMICE)

Beatriz Sanchon-Lopez and Jeremy R. Everett*

Medway Metabonomics Research Group, University of Greenwich, Chatham Maritime, Kent ME4 4TB, United Kingdom

Supporting Information

ABSTRACT: A new, simple-to-implement and quantitative approach to assessing the confidence in NMR-based identification of known metabolites is introduced. The approach is based on a topological analysis of metabolite identification information available from NMR spectroscopy studies and is a development of the metabolite identification carbon efficiency (MICE) method. New topological metabolite identification indices are introduced, analyzed, and proposed for general use, including topological metabolite identification carbon efficiency (tMICE). Because known metabolite identification is one of the key bottlenecks in either NMR-spectroscopy- or mass-spectrometry-based metabonomics/metabolomics studies, and given the fact that there is no current consensus on how to assess metabolite identification confidence, it is hoped that these new approaches and the topological indices will find utility.



KEYWORDS: *metabonomics, metabolomics, metabolic profiling, NMR spectroscopy, metabolite topology, topological metabolite identification carbon efficiency (tMICE), topological metabolite identification nitrogen and carbon efficiency (tMINCE)*

1. INTRODUCTION

Metabolic profiling of animal and human biofluids and tissues is emerging as a key technology in biology and especially in medicine,[1] where it can be used in either a diagnostic or prognostic mode. These metabonomics[2,3] or metabolomics[4] studies are typically executed with NMR-spectroscopy- or mass-spectrometry-based technologies for metabolite identification in biofluids, cell extracts, or tissue samples. There are many steps in a metabonomics experiment, and most of these steps have well-described protocols for NMR-[5−9] or mass-spectrometry (MS)-based[10−14] approaches and well-accepted statistical procedures[15−20] for their analysis, with the significant exception of known metabolite identification. This remains a problematic step for both NMR-spectroscopy-[21−25] or mass-spectrometry MS-based[10,26,27] metabonomics. The problem is essentially one of complexity and diversity. In contrast with the 4 different bases in the nucleotides of DNA and the 20 natural amino acids in protein structures, there are thousands of structurally diverse, metabolites in biofluids,[23] and their identification is not straightforward by NMR or MS.

New approaches to NMR- and MS-based known metabolite identification are emerging,[25,28] including new database and chemical treatment approaches. The new database approaches include: (i) a new isomer-specific database, $^{1}H(^{13}C)$-TOCCATA, for the identification of metabolites from TOCSY and natural abundance HSQC-TOCSY spectra,[29] (ii) a new, unified and isomer-specific, database interrogation method that provides improved HSQC-based metabolite identification performance (COLMAR),[30] and (iii) SpinCouple,[31] a new database for the analysis of the J-resolved spectra of metabolite mixtures, which contains a much larger number of spectra of metabolite standards than the similar Birmingham Metabolite Library (BML).[32] The isomer-specific features of $^{1}H(^{13}C)$-TOCCATA and COLMAR are useful because they overcome the problem of failure to identify a metabolite if only the peaks from the high abundance isomer are observed, leading to low scores in overall metabolite matching algorithms because the low abundance isomer peaks are not scored. In addition, COLMAR contains useful confidence terms including a peak matching ratio and a peak uniqueness statistic that decrease false-positives. To provide an orthogonal method for metabolite confirmation, Bingol and Brüschweiler[33] developed a method called NMR/MS Translator that confirms metabolite identifications from the HSQC COLMAR database by predicting the MS adduct ions that should be observed from the metabolites.

Metabolite identification is actually a problem in two distinct categories: first, the de novo structure elucidation of novel metabolites identified for the first time, and second, the structure confirmation of known metabolites that have been identified and characterized previously. To elucidate the structure of novel metabolites, it is generally accepted that the rigorous processes

Received: July 7, 2016

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

used in the field of natural product structure elucidation are required.[34] This would typically involve the isolation of the novel metabolite from the biological matrix and its purification ahead of full molecular structure elucidation by a panoply of spectroscopic techniques. See Supplementary Table S1 for a list of the structural features that can be revealed by these different technologies.

By contrast, the structure confirmation of the identities of known metabolites is generally accepted to require a less rigorous process because the structure has already been elucidated and spectroscopic data on the pure metabolite may be available. However, there is little consensus on what the identification process for known metabolites should be. To address this problem, the Metabolomics Standards Initiative (MSI)[35] set up a Chemical Analysis Working Group (CAWG) in 2007.[36] This group developed a four-level classification scheme for the identification of known metabolites: (1) Identified Compounds, (2) Putatively Annotated Compounds, (3) Putatively Characterized Compound Classes, and (4) Unknown Compounds.

To reach Level 1, Identified Compounds, the following requirements were stated: "A minimum of two independent and orthogonal data relative to an *authentic* compound analyzed under *identical* experimental conditions are proposed as necessary to validate non-novel metabolite identifications (e.g. retention time/index and mass spectrum, retention time and NMR spectrum, accurate mass and tandem MS, accurate mass and isotope pattern, and full ^{1}H or ^{13}C NMR, 2-D NMR spectra). The use of literature values reported for authentic samples by other laboratories is generally believed to be insufficient to validate a confident and rigorous identification. The use of literature or external laboratory data results in level 2 identifications.[36] Thus, the CAWG stipulated that any known metabolite identification made by reference to an authentic standard in a database such as the Human Metabolome Database (HMDB),[37] the BioMagResBank (BMRB),[38] and the Birmingham Metabolite Library (BML)[32] or to a literature report should be downgraded to Level 2, Putatively Annotated Compounds, whereas, if that same metabolite had been identified by reference to an authentic standard, actually in the laboratory of the investigator, it would be classed as Level 1, Identified Compound.

The original CAWG recommendations have not been widely adopted,[39] and recently new outline proposals have emerged[40−43] for improvements to the original four-level metabolite identification classification system. The proposals were either to increase the refinement of the four-level system or to introduce some sort of scoring method for the acquisition of certain sorts of data such as a 2D NMR. However, with the exception of some proposals for accurate mass and retention time fits,[42] no proposals were made as to how a match of the experimental data with the standard data should be assessed. In addition, a call to the community was made[41] for engagement with this problem. In response to this call, new quantitative proposals for Level 2 metabolite annotation confidence using LC−MS methods recently emerged from the group of Daly.[44] In addition, our group proposed some new, quantitative approaches to known metabolite identification confidence for NMR-based metabonomics studies including metabolite identification carbon efficiency (MICE),[45] drawing on an approach from the drug discovery field known as ligand efficiency.[46] The MICE methodology simply counts the number of pieces of metabolite identification information (MII) obtained for a metabolite in NMR-based metabolite identification (proton chemical shifts, signal multiplicities, coupling constants ($^{2}J_{H,H}$ and $^{3}J_{H,H}$), COSY connectivities, presence of second-order spin system, HSQC cross-peaks) and divides this sum total by the number of carbon atoms in the metabolite. The MICE metric is thus a measure of the amount of identification information obtained relative to the size of the metabolite, as judged by the number of carbon atoms it possesses. The following guidelines were proposed for a metabolite to be considered confidently identified:

(1) MICE value $\geq$1.0.

(2) Experimental data are a good fit to either authentic reference standard data or literature or database values (differences between the experimental and the expected chemical shifts within $\pm$0.03 ppm for ^{1}H and $\pm$0.5 ppm for ^{13}C and homonuclear coupling constants within $\pm$0.2 Hz).

(3) Experimental data provide good "coverage" across all parts of the molecular structure of the metabolite.

(4) Signal-to-noise ratio and resolution (actual and digital) in the spectra should be sufficient to measure the signal features with confidence.

(5) Care should be applied when assigning signals in crowded spectral regions.

(6) HSQC data are important in metabolite identification, as they provide an excellent orthogonal data source via the ^{13}C NMR chemical shift, which is much more sensitive to environment that the proton chemical shift.

(7) HMBC data should be used wherever possible to corroborate identifications that are uncertain or those that are critical for the biological interpretation of the experiment, that is, those metabolites that may be biomarkers.

This current work now reports on a new topological approach to NMR-based, known metabolite identification, which further develops the concepts introduced in the original MICE work.[45] The MICE and related analyses were novel, simple, and quantitative but had the disadvantage that a judgment had to be made of how well the MII measured represented the entire molecular structure (point 3 in the list above). This new work overcomes that issue, again, in a simple quantitative fashion by measuring the MII obtained by NMR spectroscopic methods for each separate molecular topology element (MTE) of the metabolite structure.

Molecular frameworks and topologies can be computed in a number of different ways. In the original molecular framework analysis by Bemis and Murcko,[47] molecules without rings were classed as not having a framework. This causes issues for the analysis of endogenous metabolites, where large numbers have no rings. For example, in a recent comparative analysis of the topologies of human metabolites relative to drugs, natural products, and other molecules, only one-third of the 6237 human metabolites identified in the HMDB were classed as having a framework; that is, the majority of the human metabolites had a structure that comprised one or more chains of atoms.[48] This methodology is therefore unsatisfactory: Any topological analysis of metabolites must address the particular structural features that they possess. To avoid this issue, the analysis reported here pragmatically classifies the metabolites under study as possessing one or more MTEs of just two types, rings and chains, each containing protonated and nonprotonated carbon atoms and slow-exchanging amide protons, as these are the elements that give rise to signals in ^{1}H NMR spectroscopy-based metabonomics experiments.

Metabolites may possess a single MTE, being a chain or ring, or may possess multiple MTEs separated by, for example, heteroatoms without slow-exchanging hydrogens, by quaternary carbons, or by methine (CH) carbon branching points in the structure. Simplicity, manageability, and applicability to NMR-based metabolite

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

profiling have been taken into account, as well as the possibility for automation, in the design of the entire analysis.

Topological metabolite identification carbon efficiency (tMICE) is shown to provide a simple, robust, and quantitative measure of the confidence of NMR-based identification of known metabolites in complex biological matrices.

2. MATERIAL AND METHODS

2.1. Biological Samples, Metabolites, and NMR Spectroscopy

The 100 metabolites included in this study were recently identified from ethically approved studies of the proton NMR spectra of the urine of male, wildtype, and flavin monooxygenase 5 knockout C57BL/6 mice[49] and of Italian Type 2 diabetic patients (Supplementary Table S2). This metabolite set is based on that recently reported, and the sample preparation details and NMR spectroscopy data acquisition parameters are also as previously reported.[45] The original set of 75 metabolites was augmented by additional metabolites identified in the past 12 months. The methods of identification of the metabolites were as previously reported.[45]

2.2. Analysis of Metabolite Features

The following information was abstracted for each metabolite and is provided in Supplementary Table S2:

(1) metabolite class
(2) metabolite common name
(3) metabolite IUPAC name
(4) HMDB code
(5) number of hydrogen atoms
(6) number of carbon atoms
(7) number of oxygen atoms
(8) number of nitrogen atoms
(9) number of sulfur atoms
(10) total number of heavy atoms (non-hydrogen atoms)
(11) nominal molecular mass in Daltons
(12) a flag for molecular symmetry: 1 = some element of symmetry is present in metabolite; 0 = no symmetry element in metabolite
(13) a flag for chirality: 1 = ≥1 chiral centers; 0 = 0 chiral centers in metabolite

2.3. Definition of Molecular Topology Elements and Atom Numbering

MarvinSketch was used for drawing, displaying, and characterizing chemical structures, and topology elements, Marvin 6.1.1, 2013, ChemAxon (http://www.chemaxon.com).

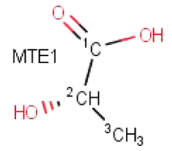
MTEs were analyzed by hand for all of the metabolites in the study, and a full list of the metabolites, their assigned MTEs, and their MII is provided in Supplementary Table S3. The following methodology and logic was applied:

(1) MTEs are defined to be of two basic types: rings of atoms (code 0) and chains of atoms (code 1).

(2) For simplicity and practicality of application, ring structures are treated as singular topology elements.

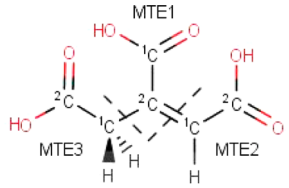
(3) The first molecular topology element, MTE1, will be the highest priority chain or ring according to IUPAC[50] functional group and Cahn–Ingold–Prelog[51] (CIP) priorities, the latter determining priorities based on the atomic number of the substituent (S > O > N > C > H, etc.) and on bond multiplicities with, for example, C=O treated as C(O)–O; that is, a carbon with a double bond to oxygen is treated as if it has two single bonds to two separate oxygen atoms.

(4) For molecules containing chains of atoms: (a) Breaks between topology elements will occur *after* quaternary carbons or heteroatoms without nonexchanging hydrogens, as these atoms interrupt the proton-to-proton connectivity information available from experiments such as 2D ^{1}H COSY NMR.[52] (b) If the start point of any MTE is a quaternary carbon it will not break; for example, there is not a break after the starting carboxylic acid carbon (C1) in (*S*)-lactic acid. The single, chain-type MTE1 goes from the C1 carboxylic acid carbon through C2 to chain termination at the methyl group, C3.



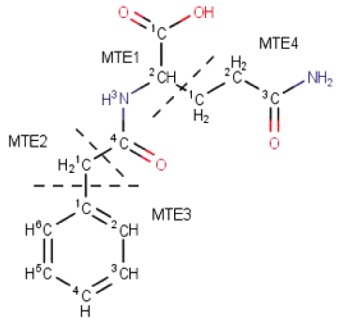
(S)-lactic acid

(c) The numbering of any subsequent MTE will start at the first atom after the break point: The breaking quaternary carbon or heteroatom is included in the preceding MTE. (d) The number of branches after a quaternary carbon will generally equal the number of atoms bonded to that quaternary carbon minus 1, and the priority for the next MTE will be given again by CIP rules. For example, the start point for MTE1 in *cis*-aconitic acid is the carboxylic acid carbon located on the central sp^2 carbon. This carboxylic acid carbon is followed by an sp^2 quaternary carbon with two branches, which go first to the =CH–COOH (MTE2) and then second to the –CH$_2$–COOH (MTE3) groups by CIP priority rules. The boundaries between MTEs are shown by dashed lines in the structures.



cis-aconitic acid

(e) If there is a branching point that is a methine carbon (CH), the MTE will move across the branch point according to CIP priority rules and will continue until chain termination or arrival at a break point. If the chain terminates, the next MTE will then be decided by CIP rules. On the contrary, if one or more new chains continue after a break point, then a structural path continuation rule applies and the subsequent MTE will carry on the same path at that point (see 5 below). For example, phenylacetylglutamine branches at the –CH– (C2) of MTE1 (HOOC–CH–NH–C=O) and continues by priority via N3 to the amide carbonyl carbon break point (C4).

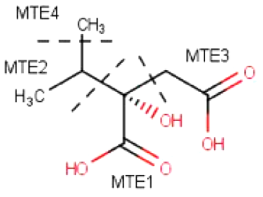


phenylacetylglutamine

The single methylene unit then becomes MTE2 as it continues the chain (rule 5 below) and coincidentally, it is higher priority (due to

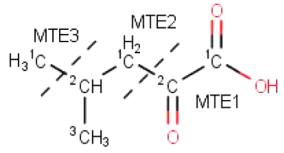
DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX–XXX

its phenyl substituent) than the alternative methylene group at the branching CH: The phenyl moiety becomes MTE3, and MTE4 is the three carbon amide chain at the methine branching point in MTE1. Another differing example occurs in (2*S*)-isopropylmalic acid, which branches at the −CH− (C1) of MTE2 (−CH−CH_3), so the − CH_2−COOH becomes MTE3 by CIP priority and then the methyl group at the branching point becomes MTE4, as this is a chain termination rather than a chain continuation example.



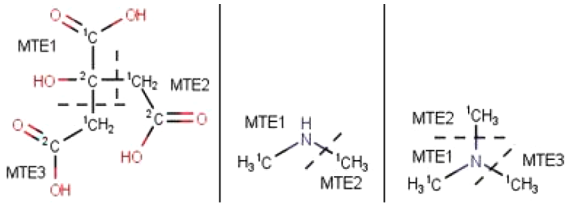
(2S)-isopropylmalic acid

(f) For metabolites with equivalent groups after a methine branching point (CH), the break will be randomly to one side or the other; for example, ketoleucine has two equivalent methyl groups: C3 in MTE2 and C1 in MTE3.



ketoleucine

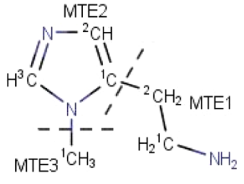
(g) For equivalent carbons allocated in different MTEs, all are registered and analyzed, for example, in ketoleucine above and in citric acid, dimethylamine, and trimethylamine, as it is critical that the topological analysis reflects the totality of the metabolite structure, including the possession of equivalent features.



MTE analysis for citric acid, dimethylamine and trimethylamine

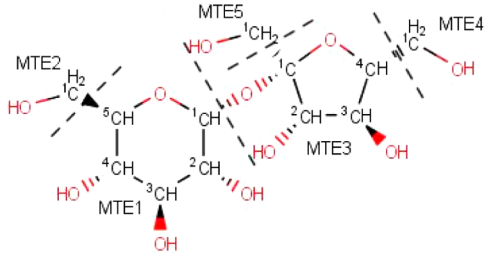
(5) Structural path continuation rule: In addition to the above rules regarding branching at CH and quaternary carbons, if after a break point there is a choice between continuing along that structural path or moving back to another structural element, continuing along the structural path will take priority (even if this breaks CIP priorities) to pragmatically align the MTE order with the metabolite structure as closely as possible. This rule applies to all metabolites, no matter of what structural type.

(6) For metabolites with a combination of rings and chains, the first molecular topology element, MTE1, will be also defined by IUPAC functional group priorities. (a) For example, MTE1 in 3-methylhistamine is the aminoethyl chain, MTE2 is the imidazole ring, and MTE3 is the methyl group.



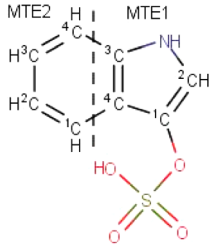
3-methylhistamine

(b) Ring priorities are decided by the number of heteroatoms of any type and if the same number, by ring size: A complex example is sucrose, whose six-membered ring (oxane) is MTE1; then, MTE2 is given by the continuation path rule to the hydroxymethyl at C5. MTE3 then corresponds to the five-membered ring (oxolane), MTE4 to the hydroxymethyl at C4, again by continuation path, and MTE5 is the remaining hydroxymethyl at C1.



sucrose

(7) For fused rings, carbons at the fusion site will be attached to the ring of greatest priority (see rule 6b), and they only will be counted once; for example, the first MTE in indoxyl-3-sulfate is the pyrrole ring (highest priority). MTE2 then corresponds to the remaining four aromatic CHs of the phenyl ring, excluding the carbons at the fusion site. The sulfate group has no observable, slow-exchanging protons and is not counted.



indoxyl-3-sulphate

2.4. Determination of the Metabolite Identification Information for each Molecular Topology Element

Once the molecular topology elements (MTEs) were defined, the experimental metabolite identification (ID) information for each metabolite was extracted from analyses of the NMR spectra of the mouse and diabetic patient urine. In parallel, the theoretical information that could possibly be obtained was calculated. Supplementary Table S2 summarizes both the theoretical and experimental information that was measured for all MTEs in all 100 metabolites.

The MII for each MTE consists of the following information, and for points 4 to 8 both actual experimental and then theoretical values are calculated:

(1) Type of MTE, coded 1 for chains and 0 for rings.
(2) Total number of carbon atoms in the MTE.
(3) Total number of nonexchanging NH groups in the MTE.
(4) MTE metabolite ID information at HSQC level,[45] which equals the sum total of the number of bits of the following information, where present and observed: (a) proton chemical shifts for each protonated carbon and nonexchanging amide NH; (b) multiplicity for each of these signals; and (c) coupling constants ($^2J_{HH}$ and $^3J_{HH}$) for each signal (note that this is different from the original MICE methodology where $^nJ_{HH}$ values were only counted once: this was seen as too conservative as each is an independent measurement). Longer-range couplings were not counted as the methodology was designed to be implemented for NMR spectra with nominal resolution: (d) presence

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

of second-order spin system (flag = 1 if there are additional lines present in the spectrum, not anticipated by a first-order spectral analysis, otherwise flag = 0); (e) *intra*-MTE COSY links via $^{2}J_{HH}$ and $^{3}J_{HH}$ between hydrogens (only counted once); (f) *inter*-MTE COSY links via $^{3}J_{HH}$ between hydrogens (only counted once and associated with the first MTE connected); and (g) HSQC cross-peaks for each protonated carbon atom (counted twice if measured separately at each of two nonequivalent hydrogens of a methylene group).

(5) HSQC-level tMICE values that equal the sum total of the information in each MTE (see point 4 above) divided by the total number of carbon atoms.

(6) HMBC connectivity information is additionally provided for metabolites, which enables the gathering of more information about information-poor MTEs. This is especially useful in symmetrical metabolites where there can be a paucity of information. The HMBC experiment can be enabling here by ^{13}C isotopomeric raising of the degeneracy within and between symmetrical MTEs. For example, even though chemically and magnetically equivalent by symmetry, HMBC cross-peaks can be observed between the methyl groups in trimethylamine due to isotopomeric raising of the degeneracy of the methyl groups, as the HMBC signals are observed via $^{3}J_{C,H}$ from the species $H_3{}^{13}C-N-{}^{12}CH_3$.

(7) HMBC connectivities were counted in two different ways: (1) HMBC connectivities inside an MTE were counted once only and (2) HMBC connectivities between MTEs were counted twice, at both the carbon-13 and proton that are connected. This is in contrast with the original MICE analysis where HMBC connectivities were counted only once. Inside an MTE, the HMBC connectivity is often between hydrogens and carbon-13 nuclei whose chemical shifts are known from HSQC-level analyses. Thus, the only new piece of information is the connection between two known signals: Therefore, we count one bit. However, inter-MTE HMBC data often provide correlations to, or through, quaternary carbons, and when it does so, it provides new information in the form of: (1) the ^{13}C NMR chemical shifts of the quaternary carbons or (2) connections between two isolated MTEs. Therefore, we counted two bits of information for *inter*-MTE HMBC connectivities, and the differential counting of *intra*- and *inter*-MTE HMBC data seems appropriate.

(8) HMBC-level tMICE+ values that correspond to the number of bits of MII at HSQC level plus the total number of HMBC links divided by the total number of carbon atoms (see Table 3 (Glossary)).

(9) A value calculated by dividing the actual experimental tMICE values by the corresponding theoretical values.

(10) As for the original MICE analysis,[45] experimental MII was classed as fitting and therefore counted when differences between the experimental and the equivalent authentic reference or literature or database chemical shifts were within ±0.03 ppm for ^{1}H and ±0.5 ppm for ^{13}C and when homonuclear coupling constants were within ±0.2 Hz.

(11) For metabolites that are symmetrical within an MTE, identification information at the HSQC level is introduced only once; for example, succinic acid has two methylene carbon atoms that are equivalent due to symmetry, but the information is only recorded once, just as it is observed.

(12) Long-range H−H COSY information is excluded, as the analysis is intended to be applicable to 2D NMR information acquired with nominal resolution and sensitivity.

2.5. Determination of the Overall Metabolite Identification Information for Each Metabolite

Having computed the metabolite ID information available experimentally and theoretically for each MTE separately in each metabolite (Section 2.3), the following information was analyzed for each metabolite as a whole:

(1) total number of carbon atoms in the metabolite

(2) total number of nonexchanging NH groups in the metabolite

(3) actual and theoretical total number of MII bits at HSQC level for all MTEs

(4) actual and theoretical total number of inter-MTE COSY links in the metabolite

(5) actual and theoretical total number of HMBC links in the metabolite

(6) actual and theoretical MICE values for each metabolite calculated by summing the total bits of information at HSQC level for all MTEs of a metabolite (point 3 above) and then dividing by the total number of carbon atoms (point 1 above)

(7) actual and theoretical MICE+ values for each metabolite were calculated by summing the total bits of information at HMBC level for all MTEs of a metabolite (point 3 plus 5 above) and then dividing by the total number of carbon atoms (point 1 above)

2.6. Precision of Measurements of ^{1}H NMR Chemical Shifts in a Variety of Metabolites in Buffered Mouse Urine and Comparison of Experimental Shifts with Those of Authentic Reference Standards in the HMDB

The chemical shifts of a variety of aliphatic, olefinic, and aromatic protons in a variety of acidic, basic, and neutral metabolites were measured to determine the variability of chemical shifts in these differing environments in mouse urine from 34 individual C57BL/6 mice and their FMO5 KO counterparts[49] at weeks 15, 30, 45, and 60. The results are summarized in Table 1 and compared with the corresponding values obtained from the HMDB.[37]

2.7. Statistical Analyses of the Data

All statistical analyses were conducted in Microsoft Excel for Macintosh version 14.5.5. All errors are standard deviations. All significance testing used the two-tailed, unpaired Student's *t* test with a confidence threshold of >95% ($p < 0.05$).

3. RESULTS

3.1. Molecular Properties of the 100 Metabolites in the Set

The 100 metabolites were in 11 classes: carboxylic acids (10), hydroxycarboxylic acids (6), dicarboxylic acids (9), tricarboxylic acids (4), small alcohols (5), ketones (2), sugars and sugar acids (7), amines (14), amides and amino acids (32), nucleosides and nucleotides (10), and others (1) for a total of 100 metabolites. See the Supplementary Tables S2 and S3 for more information. The atomic composition and molecular weight properties of the 100 metabolites are shown in Table 2.

Table 2 gives basic statistics for the number of hydrogen, carbon, oxygen, nitrogen, and sulfur atoms present in the metabolite set, including the number of metabolites possessing these atoms (first row) and the statistics for the entire set in subsequent rows, plus basic statistics on the metabolites' nominal molecular mass. While all 100 metabolites studied contained hydrogen and carbon, 95 contained oxygen, and 56 contained nitrogen, only 4 of the sets contained one sulfur atom each. Further analysis (Supplementary Table S2) shows that 16 of the 100 metabolites

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 1. Average Experimental Chemical Shifts (n = 34), Standard Deviations, Maximum and Minimum Values, Range, Corresponding HMDB Values, and Differences between the HMDB and Average Experimental Values for 10 Representative Hydrogen Environments in a Variety of Acidic, Basic, and Neutral Metabolites[a]

	β-D-fucose CH_3	*cis*-aconitic acid CH	TMA	MA	TMAO	*N*-butyrylglycine CH_2 at 1.62	3-indoxyl sulfate, 7.7	citrate, 2.56	cinnamoylglycine, 6.7	hippurate ortho
average experimental shift	1.254	5.721	2.879	2.611	3.276	1.620	7.704	2.558	6.723	7.837
standard deviation	0.0010	0.0172	0.0028	0.0024	0.0028	0.0007	0.0030	0.0053	0.0060	0.0017
maximum value	1.256	5.752	2.885	2.616	3.282	1.622	7.710	2.570	6.734	7.840
minimum value	1.252	5.696	2.875	2.607	3.272	1.619	7.698	2.549	6.706	7.832
range in values	0.004	0.056	0.010	0.009	0.010	0.003	0.012	0.021	0.028	0.008
HMDB shift	1.25	5.693	2.893	2.59	3.253	1.613	7.712	2.527	6.706	7.820
difference HMDB vs experimental	0.004	0.028	0.014	0.021	0.023	0.007	0.008	0.031	0.017	0.017

[a]TMA, trimethylamine; MA, methylamine; TMAO, trimethylamine *N*-oxide. The chemical shift of the second-order ortho protons of hippuric acid was estimated at the mid-point of the complex signal. The shifts of cinnamoylglycine were obtained from an authentic Sigma reference standard, as no data were available in the HMDB entry for this metabolite.

Table 2. Number of Metabolites Containing Hydrogen, Carbon, Oxygen, Nitrogen and Sulphur Atoms and Statistical Information on the Elemental Composition[a]

	H	C	O	N	S	total heavy atmos	nominal mass in Da
metabolite count	100	100	95	56	4		
maximum	17	13	8	5	1	21	297
minimum	2	1	0	0	0	2	31
mean	8.69	5.45	3.09	0.94	0.04	9.52	138.0
median	8	5	3	1	0	9	132.0
standard deviation	3.01	2.44	1.81	1.11	0.20	3.87	54.5

[a]All data for the set of 100 metabolites in this study.

possessed slow-exchanging NH groups that could potentially give rise to additional signals and connectivities in their ^{1}H NMR spectra.

The average molecular weight of the set is 138.0 ± 54.5 Da (standard deviation), and the distribution is shown in Figure 1.

3.2. Molecular Topology Elements in the Metabolites

Topological analysis of the set of 100 metabolites showed that they possess between 1 and 5 MTEs with the distribution shown in Figure 2. 67% of the metabolites possess 1 or 2 MTEs, 21% possess 3, 10% possess 4, and only 2% possess 5 MTEs.

The 100 metabolites contained a total of 216 MTEs, of which the vast majority (168, 77%) were chains. The distribution of MTE types for MTEs 1 to 5 is shown in Figure 3.

An overview of the distribution of the topologies and their types for all 100 metabolites is given in Figure 4, in the form of a grid. The overwhelming preponderance of chain MTEs is clear from Figures 3 and 4.

3.3. Calculation of Topological Metabolite Identification Efficiency Values

The number of bits of MII in each MTE was measured by default at the HSQC level. When this number is divided by the number of carbons in the MTE, we arrive at the tMICE value for that MTE, by analogy to the previously described MICE index.[45]

The outcome of the topological analysis for MTE1 and MTE2 for the first 16 metabolites, the carboxylic acids, and hydroxycarboxylic acids, is shown in Figure 5. Each metabolite is associated with a symmetry flag and a chirality flag that is each set to either 0 (no symmetry or no chirality) or 1 (some element of symmetry or chirality, respectively). Then, each MTE is analyzed separately in terms of the number of bits of MII at the HSQC level and the number of inter-MTE COSY links. The inter-MTE COSY information is, of course, included in the HSQC-level MII but is split out separately to give a view of how much information there is at this level to link MTEs together in the same metabolite. The analysis is conducted from both information derived experimentally and that which is potentially observable theoretically. Naturally, not all of the MII that is theoretically available is actually observable due to issues of low metabolite concentrations or crowding in the real spectra. The tMICE values can be simply calculated by dividing the total number of metabolite identification bits at the HSQC level in the MTE by the number of carbon atoms in that same MTE.

3.4. Analysis of tMICE Values for All 100 Metabolites at HSQC Level

We next analyzed the tMICE values for each of the 217 MTEs in the 100 metabolites using the three-part triage shown below (Figure 6). In the subset of metabolites illustrated in Figure 5, it can be observed that there are three kinds of situation with respect to the tMICE values for the MTEs in each metabolite:

(1) It is theoretically possible to obtain sufficient MII for the MTE to generate a tMICE of ≥1.0 (a proposed cutoff threshold value for confident identification of the MTE), and this is achieved experimentally (green).

(2) It is theoretically possible to obtain a tMICE of ≥1.0 for the MTE but this is NOT achieved experimentally (red).

(3) It is theoretically impossible to obtain a tMICE of ≥1.0 (cyan in Figure 6).

135 MTEs (62%) have actual tMICE values ≥1.0 (green); 29 MTEs (13%) have actual tMICE values <1 but theoretical tMICE values ≥1.0 (red); and 52 MTEs (24%) have theoretical tMICE values <1.0 (cyan).

3.5. tMICE+ Values for all 100 Metabolites at HMBC Level

In the tMICE+ analysis at HMBC level, 168 MTEs (77%) have actual tMICE+ values ≥1.0; 40 MTEs (19%) have actual tMICE + values <1 but theoretical tMICE values ≥1.0; and only 8 MTEs (4%) have theoretical tMICE+ values <1.0.

In moving from the HSQC-level tMICE analysis to the HMBC-level tMICE+ analysis, 27 MTEs changed from cyan (theoretical tMICE < 1) to green (actual tMICE+ ≥ 1.0); 6 MTEs changed from red (actual tMICE values <1 but theoretical tMICE values ≥1.0) to green (actual tMICE+ ≥ 1.0), and 17 MTE changed from cyan

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

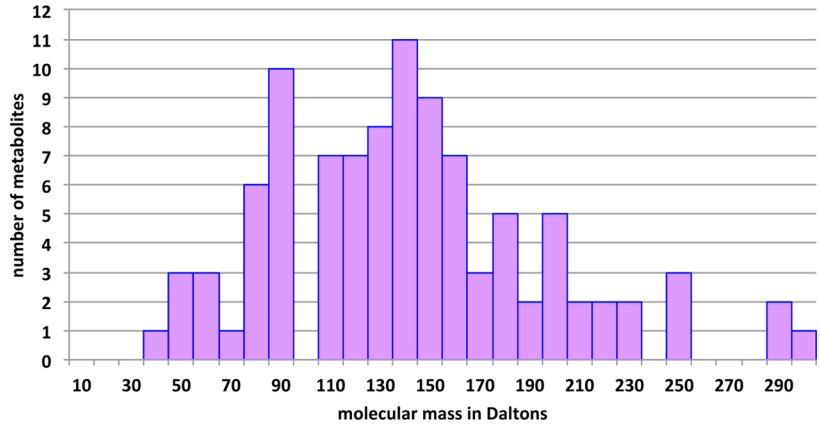


Figure 1. Histogram of the distribution of molecular weights of the 100 metabolites in the set.

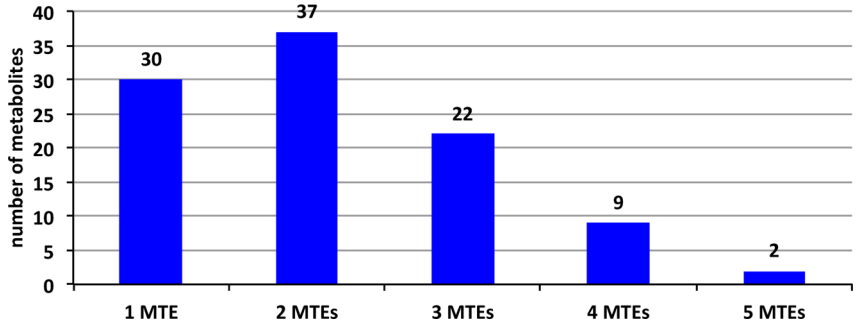


Figure 2. Chart of the number of metabolites with 1, 2, 3, 4, or 5 MTEs in the 100 metabolites studied.

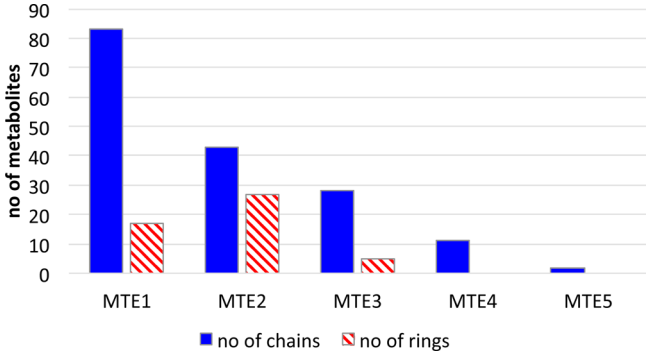


Figure 3. Chart of the distribution of chains or rings across MTEs 1 to 5 in the 100 metabolites in the current cohort.

(theoretical tMICE < 1) to red (actual tMICE+ < 1 but theoretical tMICE+ ≥ 1.0). See also Figure 7.

tMICE+ values for metabolites with some degree of symmetry (actual 2.18 ± 1.93 and theoretical 3.49 ± 2.62, n = 79) were lower than those for unsymmetrical metabolites (actual 2.68 ± 2.43, p = 0.101 and theoretical 7.20 ± 5.50, n = 137, $p = 2.30 \times 10^{-10}$) but were only statistically significantly lower for the theoretical tMICE+ values. Similarly, tMICE+ values for achiral metabolites (actual 2.37 ± 1.90 and theoretical 3.93 ± 2.31, n = 140) were lower than those for chiral metabolites (actual 2.73 ± 2.82, p = 0.318 and theoretical 9.35 ± 6.51, n = 76, $p = 4.82 \times 10^{-10}$), although again only the theoretical tMICE+ differences were statistically significant.

4. DISCUSSION

Confidence in known metabolite assignment is one of the key issues facing metabonomics/metabolomics at present. Many current studies do not report in detail their MS- or NMR-based spectroscopic analyses, let alone any assessment of the confidence ratings for the identification of important metabolites.

The Metabolomics Standards Initiative (MSI)[35] recognized this as an important issue as far back as 2007 and proposed a set of guidelines,[36] but few studies currently refer to these,[39] and improvements/amendments to the guidelines have been suggested recently.[28,41−45]

There are three key issues with the current MSI guidelines for metabolite identification in our view: (i) they are qualitative and not quantitative, referring to having data such as 2D NMR spectra; (ii) there is no guidance on how good a fit of the experimental data to reference data should be; and (iii) there is an assertion that data in the literature or in databases such as the HMDB are "generally believed insufficient to validate a confident and rigorous identification".[36] The work reported here addresses these issues. First, both the recent MICE and the new tMICE methods presented here are quantitative in basis but simple and easy to calculate. Second, clear guidance is given on the goodness of fit required for experimental NMR data to be seen as a good match for literature or reference standard data. Third, the assumption that data generated from a reference standard of a metabolite in the investigator's laboratory will be significantly different from that recorded on a different sample of that standard in the literature or in a database such as the HMDB is shown to be generally incorrect for NMR-based metabonomics (Table 1). This may be an issue and a concern for MS-based experiments such as UPLC−MS or LC−MS, where, for example, metabolite retention time could be influenced by a number of factors including the exact column type, history and age, and mass spectral intensities could be influenced by sample- and spectrometer-specific ion suppression and enhancement effects, in addition to differential

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

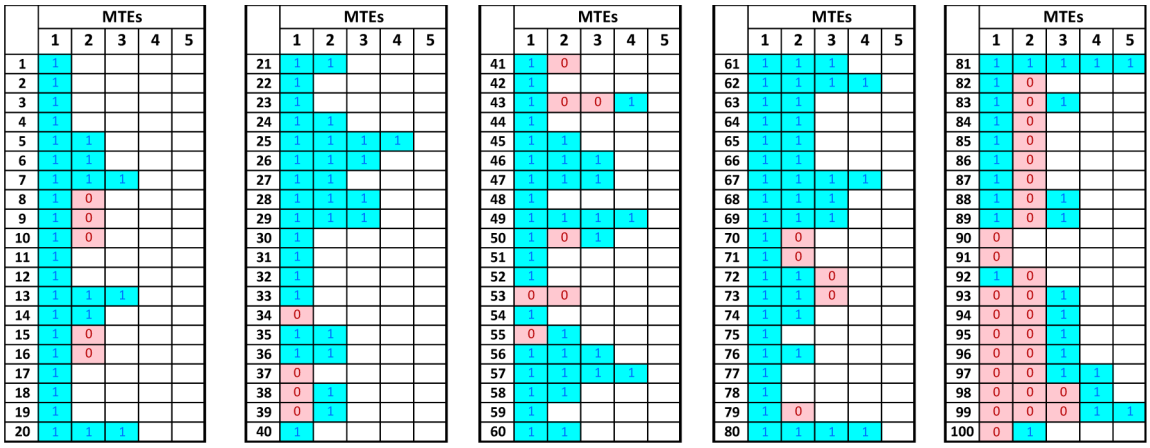


Figure 4. Simplified depiction of the MTE composition in terms of chains (blue, 1) or rings (pink, 0) for all 100 metabolites in the set.

metabolite class	COMMON NAME	HMDB code	molecular symmetry flag	chirality flag	MTE1 Type of MTE	MTE1 Total No of C atoms	MTE1 No of non-exchanging NH groups	MTE1 Actual MII HSQC level	MTE1 Actual inter-MTE COSY links	MTE1 Actual tMICE	MTE1 Theor. MII HSQC level	MTE1 Theor. inter-MTE COSY links	MTE1 Theor. tMICE	MTE2 Type of MTE	MTE2 Total No of C atoms	MTE2 No of non-exchanging NH groups	MTE2 Actual MII HSQC level	MTE2 Actual inter-MTE COSY links	MTE2 Actual tMICE	MTE2 Theor. MII HSQC level	MTE2 Theor. inter-MTE COSY links	MTE2 Theor. tMICE
carboxylic acids	formic acid	HMDB00142	0	0	1	1	0	3	0	3.0	3	0	3.0									
	acetic acid	HMDB00042	0	0	1	2	0	3	0	1.5	3	0	1.5									
	propanoic acid	HMDB00237	0	0	1	3	0	7	0	2.3	9	0	3.0									
	butyric acid	HMDB00039	0	0	1	4	0	13	0	3.3	15	0	3.8									
	isobutyric acid	HMDB01873	1	0	1	3	0	5	0	1.7	9	0	3.0	1	1	0	0	0	0.0	0	0	0.0
	isovaleric acid	HMDB00718	1	0	1	4	0	8	0	2.0	15	0	3.8	1	1	0	0	0	0.0	0	0	0.0
	ketoleucine	HMDB00695	1	0	1	2	0	0	0	0.0	0	0	0.0	1	3	0	11	0	3.7	15	0	5.0
	benzoic acid	HMDB01870	1	0	1	1	0	0	0	0.0	0	0	0.0	0	6	0	13	0	2.2	16	0	2.7
	phenylacetic acid	HMDB00209	1	0	1	2	0	2	0	1.0	3	0	1.5	0	6	0	6	0	1.0	16	0	2.7
	hydrocinnamic acid	HMDB00764	1	0	1	3	0	9	0	3.0	9	0	3.0	0	6	0	1	0	0.2	16	0	2.7
hydroxy-carboxylic acids	glycolic acid	HMDB00115	0	0	1	2	0	3	0	1.5	3	0	1.5									
	(S)-lactic acid	HMDB00190	0	1	1	3	0	8	0	2.7	9	0	3.0									
	2-hydroxyisobutyric acid	HMDB00729	1	0	1	2	0	0	0	0.0	0	0	0.0	1	1	0	3	0	3.0	3	0	3.0
	(S)-3-hydroxyisobutyric acid	R isomer is: HM	0	1	1	3	0	2	1	0.7	20	1	6.7	1	1	0	4	1	4.0	4	1	4.0
	4-hydroxybenzoic acid	HMDB00500	1	0	1	1	0	0	0	0.0	0	0	0.0	0	6	0	8	0	1.3	10	0	1.7
	4-hydroxyphenylacetic acid	HMDB00020	1	0	1	2	0	3	0	1.5	3	0	1.5	0	6	0	8	0	1.3	10	0	1.7

Figure 5. Experimental and theoretical metabolite identification information from NMR for MTEs 1 and 2 for the first 16 metabolites in the set of 100: the carboxylic acids and hydroxycarboxylic acids. MII = metabolite identification information. The tMICE values are color coded according to magnitude from low (red) to high (green). Theor. = theoretical. See the text for description and Table 3 (Glossary) for a full explanation of all terms. The entire table is available as Supplementary Table S2.

adduct formation.[10] However, in general, these concerns will not apply to ^{1}H and ^{13}C NMR spectroscopy data. With the exception of a minority of metabolites that have spectral characteristics that are particularly sensitive to the exact sample environment, the vast majority of metabolites have NMR spectral data that match very well the corresponding information on authentic standards under equivalent conditions in databases such as the HMDB. Indeed, if this was not the case, there would be little purpose in assembling these databases in the first place. Moreover, NMR data on metabolites are inherently quantitative (given certain provisos), are generally referenced to the same or highly similar chemical shift reference standards, and are not subject to the instrument/technique-specific changes in data that can be observed in MS between different ionization methods and different detection systems. Finally, it has been known for some time that the analytical variability of NMR-based metabonomics studies is low.[53−56]

In practice, of course, some differences in chemical shifts will occur, especially for ^{1}H or ^{13}C nuclei in metabolites that are subject to tautomerism or ionization changes with alteration in solvent/medium pH. To test this variability, we measured the chemical shifts of ten different hydrogen types in a variety of acidic, basic, and neutral metabolites in 34 independent mouse urine samples (Table 1). The largest range of shift values observed for one hydrogen environment was that for the olefinic CH proton of *cis*-aconitic acid, which showed a value of 5.721 ± 0.017 ppm (mean ± standard deviation) and a range of 0.056 ppm. Surprisingly, these values were greater than those for the low-frequency methylene proton signal from citrate: 2.558 ± 0.005 ppm, with a range of 0.021 ppm. As expected, the chemical shifts of the methylene protons in *N*-butyrylglycine at 1.620 ± 0.001 ppm with a range of only 0.003 ppm were almost invariant across all of the samples. The shifts of other metabolites such as trimethylamine were intermediate in variability at 2.879 ± 0.003 ppm, range 0.010 ppm. Nevertheless, the data showed that (assuming normal distribution) even for the sensitive *cis*-aconitic acid olefinic proton shifts >95% of values are expected to occur in a range of ±0.034 ppm (two times the standard deviation) around the mean, and all 34 experimental values do. For the intermediate case of trimethylamine, >95% of values would be expected to occur in a range of ±0.006 ppm around the mean, and all 34 values do.

The experimental chemical shift values were then compared with the corresponding HMDB values (or in the single case of cinnamoylglycine, an actual reference standard run on our spectrometers, as there are no data in HMDB). For the 10 hydrogen environments studied, the differences between the mean experimental mouse values and the HMDB[37]/reference values (Table 1)

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

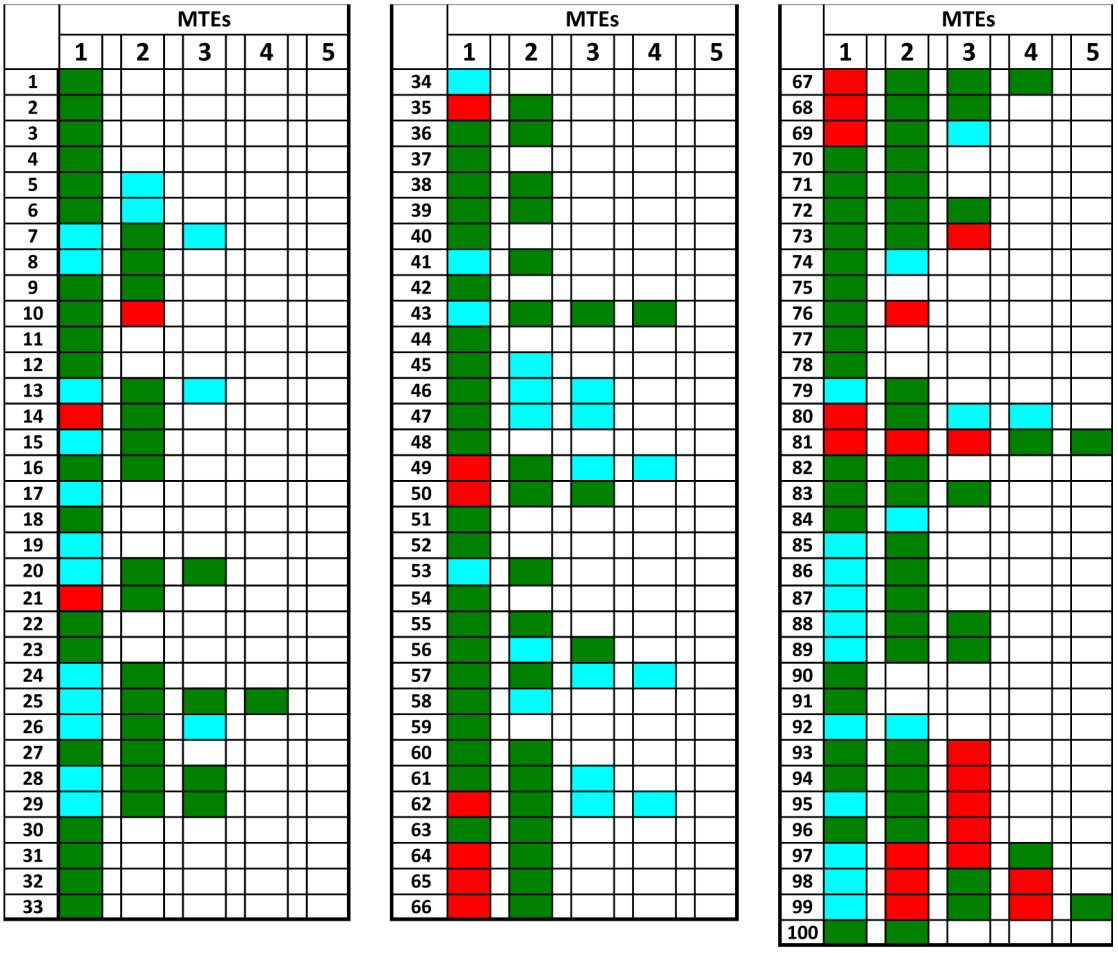


Figure 6. Analysis of the tMICE value for each MTE in each of the 100 metabolites at the HSQC level. Green: actual tMICE value ≥1.0; red: tMICE actual <1.0 but tMICE theoretical ≥1.0; cyan: tMICE theoretical <1.

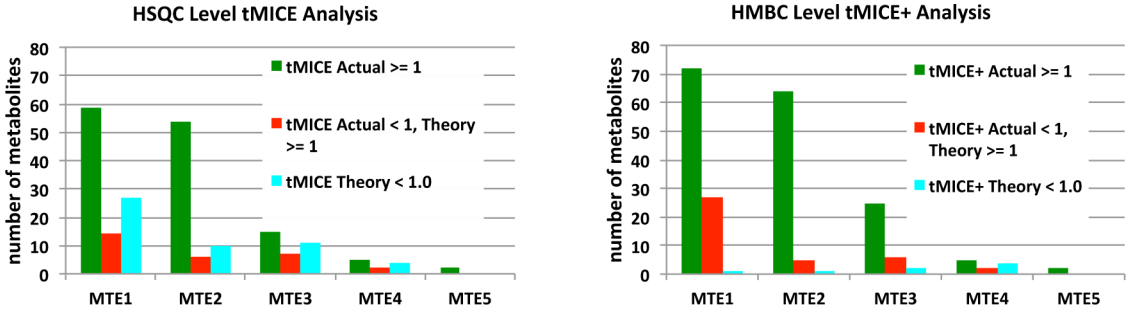


Figure 7. Left: Histogram of the distribution of HSQC-level tMICE values in MTEs: Green: actual tMICE value ≥1.0; red: tMICE actual <1.0 but tMICE theoretical ≥1.0; cyan: tMICE theoretical <1.0. Right: Corresponding HMBC-level tMICE+ analyses with analogous color coding.

varied from a low of 0.004 ppm (methyl protons in β-D-fucose) to a high of 0.031 ppm for the high-frequency methylene proton in citric acid, with an average difference of 0.017 ± 0.009 ppm (standard deviation). Our proposal that comparisons of experimental data can be confidently made with reference standard data from databases such as the HMDB rather than generating new NMR data on an authentic reference standard on the same spectrometer seems valid. In addition, the guideline of matching proton chemical shifts between experimental data and database or literature values of ±0.03 ppm also seems reasonable.

The group of 100 metabolites studied here is typical of those identified by NMR spectroscopy in metabonomics studies. These 100 metabolites have a molecular weight distribution as follows: 138.0 ± 54.5 Da in this set versus 126.7 ± 46.6 Da for the set of 75 subjected to MICE analysis.[45] Correspondingly, the number of carbon atoms in the metabolites in this set is 5.5 ± 2.4 versus 4.9 ± 2.2 atoms previously (all figures means ± standard deviations). These changes are due to the fact that further work had identified metabolites that were less obvious in the year since the previous analysis was completed. These less obvious metabolites generally had larger and more complex structures with increased molecular weight.

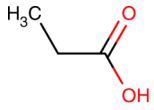
The topological approach to metabolite identification introduced here is a natural approach for NMR spectroscopy. Networks of proton-to-proton connectivity that are discovered by

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

methods such as 2D ^{1}H COSY NMR are frequently broken or interrupted by so-called "spectroscopically silent centers" such as heteroatoms bearing no slow-exchanging protons or quaternary carbons. The topological analysis defines these "spectroscopically silent centers" as two of the four types of break between the MTEs in the structure of the metabolite, with the other two breaks between MTEs being ring junctions and branching methine carbons. Thus, in this analysis, there tends to be a natural alignment between the topology elements in the metabolites and subnetworks of proton−proton connectivity derived from the NMR spectra.

The basic tMICE approach measures the number of bits of MII derived from NMR spectra from each MTE at a level up to 2D ^{1}H, ^{13}C HSQC, that is, information gathered from 1D ^{1}H NMR, 2D ^{1}H J-resolved NMR, 2D ^{1}H COSY NMR, and 2D ^{1}H, ^{13}C HSQC experiments. These bits of information include the following: (1) the number of proton chemical shifts for each protonated carbon and nonexchanging amide NH; (2) the number of signal multiplicities identified (usually from 1D ^{1}H NMR or 2D ^{1}H J-resolved NMR); (3) the number of coupling constants measured ($^{2}J_{HH}$ and $^{3}J_{HH}$) for each signal (note that this is different from the original MICE methodology where $^{n}J_{HH}$ values were only counted once: this was seen as too conservative); (4) the presence of second-order spin system (flag = 1 if there are additional lines present in the spectrum, not anticipated by a first-order spectral analysis, otherwise flag = 0; the bit is only counted if it is set to 1); (5) the number of intra-MTE COSY links via $^{2}J_{HH}$ and $^{3}J_{HH}$ between hydrogens (only counted once) and the number of inter-MTE COSY links via $^{3}J_{HH}$ between hydrogens (only counted once and associated with the first MTE connected); and (6) the number of HSQC cross-peaks for each protonated carbon atom (counted twice if measured separately at nonequivalent hydrogens of a methylene group). The sum total number of bits of information for each MTE is then divided by the number of carbon atoms in that MTE to give the tMICE value.

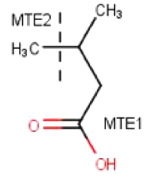
For example, simple metabolites like propanoic acid just have one chain-class MTE from C1═O to C3H$_3$.



propanoic acid

In this case, a total of seven bits of MII were experimentally observed (see Figure 5): two ^{1}H chemical shifts (1.061, C3H$_3$ and 2.190 ppm, C2H$_2$), two ^{1}H signal multiplicities (triplet and quartet), two $^{3}J_{HH}$ coupling constants (7.7 and 7.7 Hz), and a COSY between the signals for the methyl and methylene groups, all matching the values given for the authentic metabolite HMDB00237 to within ±0.03 ppm and ±0.2 Hz for ^{1}H chemical shifts and coupling constants, respectively (1.04 and 2.17 ppm, 7.7 Hz, all for a 10 mM sample at pH 7.0 in H$_2$O referenced to DSS, accessed from HMDB on April 25, 2016).[37] With three carbons in the single MTE for propanoic acid, this gives it a tMICE value of 7/3 = 2.3. Theoretically, at an HSQC level, two ^{13}C chemical shifts could have been observed, but these were below the sensitivity of our experiment and were not observed. The theoretical tMICE value is 9/3 = 3.0.

Slightly more complex metabolites like isovaleric acid were characterized as follows.



isovaleric acid

In MTE1 (four carbons) of isovaleric acid, a total of eight bits of information were experimentally obtained at an HSQC level, including two ^{1}H shifts, two multiplicities, two coupling constants, one ^{13}C shift, and one COSY connectivity (0.916, d, 6.6 Hz, 24.8 ppm with COSY to 1.958, multiplet and 2.062, d, 7.4 Hz), matching that of the authentic metabolite HMDB00718 (0.90, d, 6.6 Hz, 24.7; 1.94, triplet of septet (HMDB erroneously has doublet of quartet), 7.9, 6.6 Hz and 2.045, d, 7.5 Hz (HMDB erroneously gives 0.5 Hz?), all for pH 7.0 in H$_2$O referenced to DSS (n.b. the raw free induction decay data files associated with this entry in the HMDB are for another unrelated metabolite) accessed on April 25, 2016).[37] Note that the signal for the methine proton H3 is a very weak 1-proton triplet of septets and was only observed indirectly in the COSY and is therefore not included in the chemical shift or multiplicity counts. For MTE2, the symmetrically equivalent methyl group, no information was obtained and, at an HSQC level, no MII can be obtained because of equivalence. The tMICE values are thus 2.0 (8/4) and 0 for MTE1 and MTE2, respectively. The errors found in the HMDB data analysis do highlight the importance of checking database entries for the quality of the sample, the spectrum, and the data analysis.

The analysis of isovaleric acid highlights the importance of the topological approach to known metabolite identification confidence. While the main portion of the metabolite is well identified, there is no information available on the second methyl group in MTE2 due to molecular symmetry and equivalence.

The HSQC-level tMICE information was analyzed separately for different overall classes of metabolite. tMICE values for metabolites with some degree of symmetry (actual 1.26 ± 1.30 and theoretical 1.87 ± 1.72, n = 79) were statistically significantly lower than those for unsymmetrical metabolites (actual 1.97 ± 1.94, p = 0.0015 and theoretical 3.88 ± 3.53, n = 137, $p = 6.45 \times 10^{-8}$). It is natural that the tMICE values of the metabolites with some element of symmetry are lower than those of metabolites without symmetry, as in the metabolites with symmetry, there will be elements of the structure for which it will not be possible to obtain independent NMR data. For example, in dimethylamine, the two methyl groups are chemically equivalent by symmetry and at an HSQC level it is not possible to get information separately from these MTEs. Even though both methyl groups contribute to the singlet signal at 2.720 ppm, the information is ascribed solely to MTE1, and the second methyl group in MTE2 is allocated no bits of MII. This is a conservative approach and reflects what is actually observed.

Similarly, tMICE values for achiral metabolites (actual 1.52 ± 1.42 and theoretical 2.08 ± 1.54, n = 140) were lower than those for chiral metabolites (actual 2.04 ± 2.23, p = 0.067 and theoretical 5.11 ± 4.24, n = 76, $p = 4.14 \times 10^{-8}$), although only the theoretical tMICE differences were statistically significant; the actual tMICE differences are just insignificant. This effect is due to the chiral center causing nonequivalence of the geminal protons on the methylene carbons in chiral metabolites. In chiral metabolites, each geminal proton typically resonates at a distinct frequency, thus doubling the amount of information available from, and to, these groups.

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

The key factor in the identification of known metabolites is whether sufficient information has been collected to be confident in the identification. The tMICE approach enables this analysis to be done quantitatively for each separate topology element in the metabolite, but the question is still, how much information is enough? Following on from the original MICE work where an overall value of MICE ≥ 1.0 was determined to be generally sufficient to confidently identify a metabolite,[45] we chose to have a tMICE value of ≥1.0 also as the cutoff between confidently and not confidently identified MTEs.

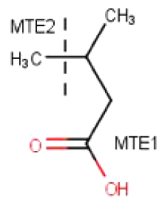
We then triaged the 100 metabolites into those where all of their MTEs were confidently identified (tMICE ≥ 1.0, green), as opposed to those metabolites that contained MTEs where the experimental tMICE values were <1.0 when the theoretical tMICE values were ≥1.0 (red) or those metabolites that contained MTEs where the theoretical tMICE value was <1.0 (cyan in Figures 6 and 7).

At a metabolite level, as opposed to an MTE level, only 42 metabolites had all of their MTEs green (tMICE ≥ 1.0, confident assignment); 24 had at least one red MTE (actual tMICE < 1.0, whereas theoretical ≥1.0), and a further 34 had at least one cyan MTE (theoretical tMICE < 1.0) with no red MTEs. Thus, 58 of the 100 metabolites had at least 1 MTE with insufficient information for confident assignment (Figure 7, all analyses for HSQC-level data).

Given that nearly a quarter of MTEs had theoretical tMICE values of <1.0 (cyan in Figure 6) and that nearly 60% of the metabolites as a whole had at least 1 MTE without sufficient information for confident assignment, the analysis was repeated at the HMBC level to generate the corresponding theoretical and experimental tMICE+ values for each MTE for comparison. The only difference between the HSQC-level tMICE value and the HMBC-level tMICE+ value for the same MTE is that the number of HMBC connectivities observed for that MTE is added to the HSQC-level MII before dividing by the number of carbon atoms in the MTE. 2D ^{1}H, ^{13}C HMBC data are very important in metabolite identification as it allows additional carbon-13 to hydrogen connectivities to be determined over two or three bonds via $^{2}J_{CH}$ and $^{3}J_{CH}$. HMBC data allows: (1) connectivities to quaternary carbons to be observed, which are otherwise generally invisible, and (2) via $^{3}J_{CH}$, connectivities through heteroatoms and quaternary carbons can be observed. These new HMBC data enable: (1) the acquisition of MII for MTEs which have significant numbers of quaternary carbons, (2) the linking of MTEs to one another, which are separated by "spectroscopically silent centers", and (3) the acquisition of information on symmetrically equivalent MTEs or parts of an MTE due to isotopomeric breaking of MTE symmetry.

Similar differences to those observed for tMICE values were seen between the tMICE+ values of both symmetric and non-symmetric metabolites and between achiral and chiral metabolites (see Results).

It is important to note that for metabolites possessing some degree of symmetry, such as isovaleric acid,



isovaleric acid

while no information could be obtained directly on the presence of the second methyl group in MTE2, due to molecular symmetry in HSQC-level experiments, in a 2D ^{1}H, ^{13}C HMBC experiment, connectivities can be seen between one methyl group and the other via the isotopomer $H_3{}^{13}C-CH-{}^{12}CH_3$, with observation at the $-{}^{12}CH_3$ group. The presence of the $H_3{}^{13}C$-isotope in one methyl group only (statistically very unlikely to observe ^{13}C isotopes in both methyl groups) means that the two methyl groups are no longer symmetrically equivalent and a true, cross-methyl HMBC connectivity is seen. Thus, this is a third important reason for the use of HMBC data in metabolite identification experiments.

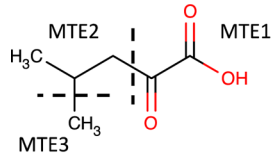
The difference between the tMICE analysis at HSQC level and the tMICE+ analysis at HMBC level is illustrated in Figure 7.

In the HSQC-level tMICE analysis, 135 MTEs (62%) have actual tMICE values ≥1.0 (green); 29 MTEs (13%) have actual tMICE values <1.0 but theoretical tMICE values ≥1.0 (red); and 52 MTEs (24%) have theoretical tMICE values <1.0 (cyan). By contrast, in the tMICE+ analysis at HMBC level, 168 MTEs (77%) have actual tMICE+ values ≥1.0; 40 MTEs (19%) have actual tMICE+ values <1.0 but theoretical tMICE values ≥1.0; and only 8 MTEs (4%) have theoretical tMICE+ values <1.0.

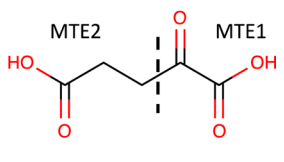
At an HMBC level, 61 metabolites now have all MTEs with sufficient identification information for confident assignment (green, cf. 42 in tMICE level), but 34 still have at least one MTE with an actual tMICE+ < 1.0, while the theoretical tMICE+ ≥ 1.0 (red, cf. 24 in tMICE analysis) and only 5 now have at least 1 MTE with a theoretical tMICE+ of <1.0 (cyan, cf. 34 in tMICE analysis), with no red MTEs.

It is important to address the issue of MTEs that do have tMICE+ < 1.0. In cases where the actual tMICE+ value is <1.0 but the theoretical value is ≥1, it is possible to look for the missing information in other samples for confirmation. In cases where the theoretical tMICE+ is <1.0, this is not possible and other approaches will be needed. We illustrate these approaches now with some examples from this work.

Ketoleucine and 2-oxoglutaric acid are examples where even though the theoretical tMICE+ was >1.0 the actual HMBC-level tMICE+ values in both cases were <1.0, and in fact no information was obtained from MTE1 at an HMBC level in either case. This is because, the first MTE in both cases is a carboxylate-keto group with two quaternary carbons joined together, resulting in no information at the HSQC level and a difficulty in obtaining HMBC-level data for samples in which metabolite concentrations are low.



ketoleucine



2-oxoglutaric acid

In these cases, we were fortunate to have access to female C57BL/6 urine samples containing higher concentrations of the metabolites (the original samples were all from male mice). For ketoleucine, female C57BL/6 mice have 2.619 (d, 7.1 Hz, CH_2), 51.1 with COSY to 2.10 (CH), TOCSY to 0.944 (CH_3), and HMBC to 24.2 and 210.8 (ketone carbonyl), thus linking MTE1 to MTE2 and providing information on MTE1. In addition, they have 0.941 (d, 6.8 Hz, CH_3), 24.5 with COSY to 2.103, TOCSY to 2.614, and HMBC to 50.7 and 24.0 (low digital resolution in

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

t1, so shift imprecise; cross-methyl connectivity), thus linking MTE 2 to the symmetrically equivalent group in MTE3 and confirming the structure. For 2-oxoglutaric acid, female C57BL/6 mice have 2.448 (t, 6.8 Hz), 33.4 with COSY to 3.013 (t, 6.9 Hz), 38.4 and HMBC from 2.444 to 184.2 and 207.6. The HMBC link from the methylene to the ketone carbon at 207.6 proves the linking of MTE1 to MTE2 and confirms the identity of the metabolite.

Trimethylamine and trimethylamine-*N*-oxide are examples of metabolites where, even at an HMBC level, it is impossible to obtain any identification information on the third MTE, which is a symmetrically equivalent methyl group. HMBC data provide information that there is a symmetrically equivalent methyl group from the cross-methyl HMBC peaks, but it cannot define that there are two of these. In these cases, the situation is made more difficult by the fact that there is so little information available: The signals of both these metabolites comprise just one singlet. Their identification is considered safe but not definitive on the basis of spectroscopic signal density and signal characteristics arguments, as follows. The information density in the ^{1}H NMR spectrum of a biofluid such as urine varies with the chemical shift and is not at a maximum in the region of the signals from trimethylamine and trimethylamine-*N*-oxide. In addition, the sharp singlet signals from these metabolites are relatively characteristic. Nevertheless, to be prudent and avoid errors, it would be recommended to confirm the identities of these metabolites by a complementary technology such as MS if they were determined to be statistically significant biomarkers in a study. From the NMR spectroscopy perspective, they are considered Putatively Annotated, Level 2 in the MSI notation. It is possible that in other cases additional information could be obtained from long-range, proton−proton coupling constants, as $^{4}J_{HH}$, $^{5}J_{HH}$, and even $^{6}J_{HH}$ couplings can be either observed directly or inferred from COSY experiments in biofluid NMR spectra.[25] In addition, information from heteronuclear couplings such as $^{2}J_{NH}$ due to ^{14}N isotopes in highly symmetrical environments[25] or $^{3}J_{PH}$ and $^{4}J_{PH}$ due to the presence of ^{31}P in nucleotides may also provide additional confirmatory information.

Other approaches to the analysis are possible. One alternative topological measure of metabolite identification confidence at the HSQC level would be the topological metabolite identification nitrogen and carbon efficiency index: tMINCE, which is equivalent to tMICE but where the number of bits of MII for an MTE, is divided by the total number of nitrogens with non-exchanging protons (NHs) plus the number of carbon atoms in the MTE. The analogous tMINCE+ index could be used for HMBC-level information. All of the information required to calculate this is provided in Supplementary Table S2. In this group of 100 metabolites, only 17 MTEs and 16 metabolites had nitrogens with slow-exchanging hydrogens attached (urea has two such MTEs). The tMINCE or tMINCE+ approaches are, in principle, more conservative than the tMICE or tMICE+ approaches due to the denominator in the equation being at least as large or larger, but in practice not one single MTE would have changed classification, that is, actual or theoretical tMINCE or tMINCE+ < 1.0 when tMICE or tMICE+ ≥ 1.0, respectively.

5. CONCLUSIONS AND FUTURE WORK

The new topological approach to metabolite identification confidence (tMICE) presented here is an improvement upon the original metabolite identification carbon efficiency (MICE) method, as it objectively monitors whether MII covers all relevant parts of the metabolite's structure. It is clear when reviewing this more stringent approach to metabolite identification confidence

Table 3. Glossary

no.	term	explanation
1	HSQC	2D, heteronuclear single quantum coherence spectroscopy, allowing correlations between protons and directly attached ^{13}C nuclei to be elucidated (2D ^{1}H, ^{13}C HSQC)
2	HMBC	2D heteronuclear multiple bond correlation spectroscopy, allowing correlations between protons and ^{13}C nuclei two to three bonds away to be elucidated (2D ^{1}H, ^{13}C HMBC)
3	MTE	molecular topology element: a chain or ring in a metabolite that forms part or all of the structure
4	MII	metabolite identification information
5	MICE	metabolite identification carbon efficiency[45] •can be defined at a number of levels •at HSQC level, equals sum total of information bits from proton chemical shifts, multiplicities, coupling constants, second-order flag, COSY connectivities, and HSQC connectivities divided by the total number of carbon atoms in the metabolite, that is, equals MII at HSQC level divided by the number of carbon atoms in the metabolite
6	MINCE	metabolite identification nitrogen and carbon efficiency •can be defined at a number of levels •at HSQC level, equals sum total of information bits from proton chemical shifts, multiplicities, coupling constants, second-order flag, COSY connectivities, and HSQC connectivities, divided by the total number of carbon atoms and nonexchanging NHs in the metabolite
7	tMICE	topological MICE •can be defined at a number of levels •at HSQC level, equals sum total of information bits from proton chemical shifts, multiplicities, coupling constants, second-order flag, intra- and inter-MTE COSY connectivities, and HSQC connectivities, that is, the MII, divided by the total number of carbon atoms in the MTE
8	tMINCE	topological MINCE •can be defined at a number of levels •at HSQC level, equals sum total of information bits from proton chemical shifts, multiplicities, coupling constants, second-order flags, intra-and inter-MTE COSY connectivities, and HSQC connectivities, divided by the total number of carbon atoms and nonexchanging NHs in the MTE
9	tMICE+	topological MICE+ •equivalent to tMICE at HMBC level, and equals sum total of information bits from proton chemical shifts, multiplicities, coupling constants, second-order flag, intra- and inter-MTE COSY connectivities, HSQC connectivities, and HMBC connectivities, divided by the total number of carbon atoms in the molecular topology element (MTE)
10	tMINCE+	topological MINCE+ •equivalent measure to tMINCE at HMBC level and equal to the total amount of information divided by the total number of carbon and nonexchanging NHs in the MTE

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

that, in many cases, HMBC-level information is important for improved confidence. This is especially the case for metabolites that (i) possess MTEs that are bounded by, or composed largely of, quaternary carbons or heteroatoms without slow-exchanging hydrogens, or (ii) possess MTEs that are rendered wholly or partly "HSQC-invisible" due to some form of symmetry, which can be broken by observation of long-range connectivities from an asymmetric ^{13}C isotopomer of the metabolite in an HMBC experiment.

The tMICE and tMICE+ methodology systematizes NMR-based known metabolite identification by (i) taking a coherent topological approach, ensuring that each element of a metabolite's structure is considered in the analysis, (ii) using a simple quantitative measure of the number of bits of MII available in each MTE, and (iii) expressing that amount of MII in ratio to the number of carbon atoms in each MTE: tMICE = (no. of MII bits)/number of carbon atoms in the MTE.

In the future, automated topological analysis of all possible metabolite structures would be advantageous, and this should be readily computed, as the rules for topology definition are quite deterministic, while aligned with the needs of NMR-based metabolite identification. In addition, the automated analysis of the theoretical tMICE and tMICE+ scores for each MTE for each metabolite should be readily computable. Metabolite identification would also be aided by precise and accurate NMR prediction programs, as it would then be possible to identify many metabolites, known and unknown by comparison with computed as opposed to experimental chemical shifts and other parameters, as experimental values are often missing.

The analysis of tMICE indices is recommended for all key biomarkers discovered in untargeted NMR-based metabonomics studies to give a measure of confidence in any biological conclusions drawn from the identification of these biomarkers. We see no reason to apply the methodology to all of the known metabolites identified and would advocate an approach that is as simple as possible to implement and use, consistent with rigor.

Known metabolite identification is a significant issue for both MS- and NMR-based metabolic profiling experiments. We hope that the tMICE and tMICE+ approaches and variants will find utility in the field. We expect that other researchers will seek to improve on these initial proposals and also develop similar approaches for MS-based metabolic profiling.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00631.

Table S1. Metabolite molecular structure information determined by different spectroscopic techniques. Table S2. Metabolite names and properties and experimental and theoretical tMICE and tMICE+ analyses. Table S3. Analysis of molecular topology elements and associated metabolite identification information, both experimental and theoretical. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +44 (0)208 331 8323. E-mail: j.r.everett@greenwich.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Professors Jeremy Nicholson, John Lindon, and Elaine Holmes for long-term collaborations on metabonomics and for access to Imperial College 600 MHz NMR facilities; Anthony Dona, Beatriz Jiminez, and Michael Kyriakides for assistance with NMR spectroscopy, Professor Elizabeth Shephard and Dr. Flora Scott for a collaboration on mouse phenotypes; Professor Stefano Balducci for a collaboration on the effects of exercise on diabetic patients; and Dr. Dorsa Varshavi for assistance with the sample preparation, spectral analysis, and metabolite identification.

■ REFERENCES

(1) *Metabolic Phenotyping in Personalized and Public Healthcare*; Nicholson, J. K., Darzi, A., Holmes, E., Lindon, J. C., Eds.; Academic Press: London, 2016.

(2) Lindon, J.; Nicholson, J.; Holmes, E.; Everett, J. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts Magn. Reson.* **2000**, *12* (5), 289−320.

(3) Lindon, J. C.; Nicholson, J. K.; Holmes, E. *The Handbook of Metabonomics and Metabolomics*; Elsevier: Oxford, U.K., 2007.

(4) Fiehn, O. Metabolomics–the link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48* (1−2), 155−71.

(5) Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.* **2007**, *2* (11), 2692−2703.

(6) Dona, A. C.; Jimenez, B.; Schaefer, H.; Humpfer, E.; Spraul, M.; Lewis, M. R.; Pearce, J. T. M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Precision High-Throughput Proton NMR Spectroscopy of Human Urine, Serum, and Plasma for Large-Scale Metabolic Phenotyping. *Anal. Chem.* **2014**, *86* (19), 9887−9894.

(7) Emwas, A.-H. M.; Salek, R. M.; Griffin, J. L.; Merzaban, J. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics* **2013**, *9* (5), 1048−1072.

(8) Pinto, J.; Domingues, M. R. M.; Galhano, E.; Pita, C.; Almeida, M. d. C.; Carreira, I. M.; Gil, A. M. Human plasma stability during handling and storage: impact on NMR metabolomics. *Analyst* **2014**, *139* (5), 1168−1177.

(9) Emwas, A.-H.; Luchinat, C.; Turano, P.; Tenori, L.; Roy, R.; Salek, R. M.; Ryan, D.; Merzaban, J. S.; Kaddurah-Daouk, R.; Zeri, A. C.; Gowda, G. A. N.; Raftery, D.; Wang, Y.; Brennan, L.; Wishart, D. S. Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review. *Metabolomics* **2015**, *11* (4), 872−894.

(10) Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.* **2010**, *2* (1−4), 23−60.

(11) Want, E. J.; Wilson, I. D.; Gika, H.; Theodoridis, G.; Plumb, R. S.; Shockcor, J.; Holmes, E.; Nicholson, J. K. Global metabolic profiling procedures for urine using UPLC-MS. *Nat. Protoc.* **2010**, *5* (6), 1005−1018.

(12) Theodoridis, G.; Gika, H. G.; Wilson, I. D. Mass Spectrometry-Based Holistic Analytical Approaches For Metabolite Profiling In Systems Biology Studies. *Mass Spectrom. Rev.* **2011**, *30* (5), 884−906.

(13) Ramautar, R.; Nevedomskaya, E.; Mayboroda, O. A.; Deelder, A. M.; Wilson, I. D.; Gika, H. G.; Theodoridis, G. A.; Somsen, G. W.; de Jong, G. J. Metabolic profiling of human urine by CE-MS using a positively charged capillary coating and comparison with UPLC-MS. *Mol. BioSyst.* **2011**, *7* (1), 194−199.

(14) Gika, H.; Theodoridis, G. Sample preparation prior to the LC-MS-based metabolomics/metabonomics of blood-derived samples. *Bioanalysis* **2011**, *3* (14), 1647−1661.

(15) Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J. K.; Lindon, J. C. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal. Chem.* **2006**, *78* (7), 2262−2267.

(16) Bylesjo, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. OPLS discriminant analysis: combining the

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

strengths of PLS-DA and SIMCA classification. *J. Chemom.* **2006**, *20* (8−10), 341−351.
(17) Izquierdo-Garcia, J. L.; Villa, P.; Kyriazis, A.; del Puerto-Nevado, L.; Perez-Rial, S.; Rodriguez, I.; Hernandez, N.; Ruiz-Cabello, J. Descriptive review of current NMR-based metabolomic data analysis packages. *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *59* (3), 263−270.
(18) Veselkov, K. A.; Vingara, L. K.; Masson, P.; Robinette, S. L.; Want, E.; Li, J. V.; Barton, R. H.; Boursier-Neyret, C.; Walther, B.; Ebbels, T. M.; Pelczer, I.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery. *Anal. Chem.* **2011**, *83* (15), 5864−5872.
(19) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M. D.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. Recursive Segment-Wise Peak Alignment of Biological H-1 NMR Spectra for Improved Metabolic Biomarker Recovery. *Anal. Chem.* **2009**, *81* (1), 56−66.
(20) Robinette, S. L.; Lindon, J. C.; Nicholson, J. K. Statistical Spectroscopic Tools for Biomarker Discovery and Systems Medicine. *Anal. Chem.* **2013**, *85* (11), 5297−5303.
(21) Fonville, J. M.; Maher, A. D.; Coen, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Evaluation of Full-Resolution J-Resolved H-1 NMR Projections of Biofluids for Metabonomics Information Retrieval and Biomarker Identification. *Anal. Chem.* **2010**, *82* (5), 1811−1821.
(22) Viant, M. R.; Ludwig, C.; Gunther, U. L.; Griffiths, W. J. 1D and 2D NMR spectroscopy: From metabolic fingerprinting to profiling. In Metabolomics, Metabonomics and Metabolite Profiling. *RSC Biomol. Sci.* **2007**, 44−70.
(23) Wishart, D. S. Advances in metabolite identification. *Bioanalysis* **2011**, *3* (15), 1769−1782.
(24) van der Hooft, J. J. J.; de Vos, R. C. H.; Ridder, L.; Vervoort, J.; Bino, R. J. Structural elucidation of low abundant metabolites in complex sample matrices. *Metabolomics* **2013**, *9* (5), 1009−1018.
(25) Dona, A. C.; Kyriakides, M.; Scott, F.; Shephard, E. A.; Varshavi, D.; Veselkov, K.; Everett, J. R. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 135−153.
(26) Watson, D. G. A rough guide to metabolite identification using high resolution liquid chromatography mass spectrometry in metabolomic profiling in metazoans. *Comput. Struct. Biotechnol. J.* **2013**, *4*, 1−10.
(27) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **2013**, *9* (1), S44−S66.
(28) Bingol, K.; Bruschweiler-Li, L.; Li, D.; Zhang, B.; Xie, M.; Brüschweiler, R. Emerging new strategies for successful metabolite identification in metabolomics. *Bioanalysis* **2016**, *8* (6), 557−573.
(29) Bingol, K.; Bruschweiler-Li, L.; Li, D.-W.; Brueschweiler, R. Customized Metabolomics Database for the Analysis of NMR H-1-H-1 TOCSY and C-13-H-1 HSQC-TOCSY Spectra of Complex Mixtures. *Anal. Chem.* **2014**, *86* (11), 5494−5501.
(30) Bingol, K.; Li, D.-W.; Bruschweiler-Li, L.; Cabrera, O. A.; Megraw, T.; Zhang, F.; Brueschweiler, R. Unified and Isomer-Specific NMR Metabolomics Database for the Accurate Analysis of C-13-H-1 HSQC Spectra. *ACS Chem. Biol.* **2015**, *10* (2), 452−459.
(31) Kikuchi, J.; Tsuboi, Y.; Komatsu, K.; Gomi, M.; Chikayama, E.; Date, Y. Spin Couple: Development of a Web Tool for Analyzing Metabolite Mixtures via Two-Dimensional J-Resolved NMR Database. *Anal. Chem.* **2016**, *88* (1), 659−665.
(32) Ludwig, C.; Easton, J. M.; Lodi, A.; Tiziani, S.; Manzoor, S. E.; Southam, A. D.; Byrne, J. J.; Bishop, L. M.; He, S.; Arvanitis, T. N.; Guenther, U. L.; Viant, M. R. Birmingham Metabolite Library: a publicly accessible database of 1-D H-1 and 2-D H-1 J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* **2012**, *8* (1), 8−18.
(33) Bingol, K.; Brueschweiler, R. NMR/MS Translator for the Enhanced Simultaneous Analysis of Metabolomics Mixtures by NMR Spectroscopy and Mass Spectrometry: Application to Human Urine. *J. Proteome Res.* **2015**, *14* (6), 2642−2648.
(34) Baker, G.; Dorgan, R.; Everett, J.; Hood, J.; Poulton, M. A Novel Series of Milbemycin Antibiotics from Streptomyces Strain-E225.2. Isolation, Characterization, Structure Elucidation and Solution Conformations. *J. Antibiot.* **1990**, *43* (9), 1069−1076.
(35) Fiehn, O.; Robertson, D.; Griffin, J.; van der Werf, M.; Nikolau, B.; Morrison, N.; Sumner, L. W.; Goodacre, R.; Hardy, N. W.; Taylor, C.; Fostel, J.; Kristal, B.; Kaddurah-Daouk, R.; Mendes, P.; van Ommen, B.; Lindon, J. C.; Sansone, S.-A. The metabolomics standards initiative (MSI). *Metabolomics* **2007**, *3* (3), 175−178.
(36) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W. M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3* (3), 211−221.
(37) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res.* **2013**, *41* (D1), D801−D807.
(38) 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.; Markley, J. L. BioMagResBank. *Nucleic Acids Res.* **2007**, *36*, D402−D408.
(39) Salek, R. M.; Steinbeck, C.; Viant, M. R.; Goodacre, R.; Dunn, W. B. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience* **2013**, *2* (1), 13−13.
(40) Jeon, J.; Kurth, D.; Hollender, J. Biotransformation Pathways of Biocides and Pharmaceuticals in Freshwater Crustaceans Based on Structure Elucidation of Metabolites Using High Resolution Mass Spectrometry. *Chem. Res. Toxicol.* **2013**, *26* (3), 313−324.
(41) 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.; Trengove, R.; Wolfender, J.-L. Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics* **2014**, *10* (3), 350−353.
(42) Sumner, L.; Lei, Z.; Nikolau, B.; Saito, K.; Roessner, U.; Trengove, R. Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics* **2014**, *10* (6), 1047−1049.
(43) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014**, *48* (4), 2097−2098.
(44) Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K. E. V.; Breitling, R. MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* **2014**, *30* (19), 2764−2771.
(45) Everett, J. R. A new paradigm for known metabolite identification in metabonomics/metabolomics: metabolite identification efficiency. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 131−44.
(46) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9* (10), 430−431.
(47) Bemis, G. W.; Murcko, M. A. The properties of known drugs 0.1. Molecular frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.
(48) Chen, H.; Engkvist, O.; Blomberg, N.; Li, J. A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *MedChemComm* **2012**, *3* (3), 312−321.
(49) Malagon, S. G. G.; Melidoni, A. N.; Hernandez, D.; Omar, B. A.; Houseman, L.; Veeravalli, S.; Scott, F.; Varshavi, D.; Everett, J.; Tsuchiya, Y.; Timms, J. F.; Phillips, I. R.; Shephard, E. A. The phenotype of a knockout mouse identifies flavin-containing monooxygenase 5 (FMO5) as a regulator of metabolic ageing. *Biochem. Pharmacol.* **2015**, *96* (3), 267−277.

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX

(50) Favre, H. A.; Powell, W. H. *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013*; Royal Society of Chemistry: Cambridge, U.K., 2013.
(51) Klein, D. R. *Organic chemistry*. John Wiley: Hoboken, N.J., 2012.
(52) Claridge, T. *High-Resolution NMR Techniques in Organic Chemistry*, 2nd ed.; Elsevier: Oxford, U.K., 2009.
(53) Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Schlotterbeck, G.; Senn, H.; Niederhauser, U.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Analytical reproducibility in H-1 NMR-based metabonomic urinalysis. *Chem. Res. Toxicol.* **2002**, *15* (11), 1380−1386.
(54) Barton, R. H.; Nicholson, J. K.; Elliott, P.; Holmes, E. High-throughput 1H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study. *Int. J. Epidemiol.* **2008**, *37* (Suppl1), i31−40.
(55) Schreier, C.; Kremer, W.; Huber, F.; Neumann, S.; Pagel, P.; Lienemann, K.; Pestel, S. Reproducibility of NMR Analysis of Urine Samples: Impact of Sample Preparation, Storage Conditions, and Animal Health Status. *BioMed Res. Int.* **2013**, *2013*, 1−19.
(56) Martin, J. C.; Maillot, M.; Mazerolles, G.; Verdu, A.; Lyan, B.; Migné, C.; Defoort, C.; Canlet, C.; Junot, C.; Guillou, C.; Manach, C.; Jabob, D.; Bouveresse, D. J.; Paris, E.; Pujos-Guillot, E.; Jourdan, F.; Giacomoni, F.; Courant, F.; Favé, G.; Le Gall, G.; Chassaigne, H.; Tabet, J. C.; Martin, J. F.; Antignac, J. P.; Shintu, L.; Defernez, M.; Philo, M.; Alexandre-Gouaubau, M. C.; Amiot-Carlin, M. J.; Bossis, M.; Triba, M. N.; Stojilkovic, N.; Banzet, N.; Molinié, R.; Bott, R.; Goulitquer, S.; Caldarelli, S.; Rutledge, D. N. Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument inter-laboratory study. *Metabolomics* **2015**, *11* (4), 807−821.

DOI: 10.1021/acs.jproteome.6b00631
J. Proteome Res. XXXX, XXX, XXX−XXX