

RESEARCH ARTICLE

Open Access



An initial investigation of accuracy required for the identification of small molecules in complex samples using quantum chemical calculated NMR chemical shifts

Yasemin Yesiltepe^{1,2}, Niranjana Govind², Thomas O. Metz¹ and Ryan S. Renslow^{1,2*} 

Abstract

The majority of primary and secondary metabolites in nature have yet to be identified, representing a major challenge for metabolomics studies that currently require reference libraries from analyses of authentic compounds. Using currently available analytical methods, complete chemical characterization of metabolomes is infeasible for both technical and economic reasons. For example, unambiguous identification of metabolites is limited by the availability of authentic chemical standards, which, for the majority of molecules, do not exist. Computationally predicted or calculated data are a viable solution to expand the currently limited metabolite reference libraries, if such methods are shown to be sufficiently accurate. For example, determining nuclear magnetic resonance (NMR) spectroscopy spectra *in silico* has shown promise in the identification and delineation of metabolite structures. Many researchers have been taking advantage of density functional theory (DFT), a computationally inexpensive yet reputable method for the prediction of carbon and proton NMR spectra of metabolites. However, such methods are expected to have some error in predicted ¹³C and ¹H NMR spectra with respect to experimentally measured values. This leads us to the question—what accuracy is required in predicted ¹³C and ¹H NMR chemical shifts for confident metabolite identification? Using the set of 11,716 small molecules found in the Human Metabolome Database (HMDB), we simulated both experimental and theoretical NMR chemical shift databases. We investigated the level of accuracy required for identification of metabolites in simulated pure and impure samples by matching predicted chemical shifts to experimental data. We found 90% or more of molecules in simulated pure samples can be successfully identified when errors of ¹H and ¹³C chemical shifts in water are below 0.6 and 7.1 ppm, respectively, and below 0.5 and 4.6 ppm in chloroform solvation, respectively. In simulated complex mixtures, as the complexity of the mixture increased, greater accuracy of the calculated chemical shifts was required, as expected. However, if the number of molecules in the mixture is known, e.g., when NMR is combined with MS and sample complexity is low, the likelihood of confident molecular identification increased by 90%.

Keywords: Metabolomics, Small molecules, NMR, DFT, Quantum chemistry

Introduction

Metabolomics and exposomics involve the large-scale study of small molecules found in biological and environmental samples, including endogenous and exogenous chemicals, and their molecular breakdown products [1–3]. For human studies, understanding the active

*Correspondence: ryan.renslow@pnnl.gov

¹The Gene and Linda Voiland School of Chemical Engineering and Bioengineering, Washington State University, Pullman, WA, USA
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

metabolic pathways and fate of exogenous chemicals is a major focus area for improving health through precision medicine, as well as an important tool for researching and understanding the state of environmental and agricultural conditions [4–7]. Biological and environmental samples typically comprise numerous molecules and are often in a complex matrix. It is not practical to develop or apply sample preparation methods for isolation of individual constituents (whether because of concentration limits, separation difficulty, or project cost limitations). The ability to comprehensively characterize such complex samples would result in significant advances in multiple scientific fields and enable currently irresolvable solutions for the understanding of metabolic pathways and biological systems such as active phenotypes/functions, industrial reactions such as they relate to (bio)fuels and high-value (bio)products, environmental processes, human actions in society, and even earth systems and climate.

Using nuclear magnetic resonance (NMR) [8–10], mass spectrometry (MS) [11–13], and other tools [14–16], a wide range of molecules have been identified and extensively documented in the literature [17–21]. Hundreds of thousands of metabolites are now known and their MS/MS or NMR data are electronically available on public and commercial chemical databases [22, 23] such as PubChem [24], Royal Society of Chemistry ChemSpider [25], ChEMBL by European Molecular Biology Laboratory [26, 27], Chemical Entities of Biological Interest (ChEBI) [28, 29], DrugBank [30, 31], Biological Magnetic Resonance Bank (BMRB) [32] and Human Metabolome Database (HMDB) [33], GDB13 [34], The Small Molecule Pathway Database (SMPDB) [35, 36], Distributed Structure-Searchable Toxicity (DSSTox) Database [37], E. coli Metabolome Database (ECMDB) [38, 39], EcoCyc E. coli Database [40], Food Component Database (FoodDB) [41], LIPID MAPS In-Silico Structure Database (LMISSD) [42], MetaCyc Metabolic Pathway Database [43], MolMall [44], Super Natural II [45], The Toxin and Toxin Target Database (T3DB) [46, 47], ToxCast [48], The Universal Natural Products Database (UNPD) [49], ZINC [50]. However, the vast majority of molecules that are found in complex biological and environmental samples are not represented in current identification libraries (across multiple analytical platforms) [51, 52]. For example, the largest mass spectral library, the Wiley Registry and NIST Libraries contain more than 1 million mass spectra [53, 54]. HMDB (ver. 4.0) describes 114,260 metabolites, and of the molecules described in HMDB, only a small portion are available for purchase as authentic reference material [55–57]. ZINC 15, a database of ~1.8 B compounds, currently has 81,519 endogenous human metabolite structures, and of these, 9490 (12%)

are immediately available for purchase [58, 59]. Furthermore, it is hypothesized that 10^{60} or more molecules are structurally feasible (for molecules < 1000 Da) [60–62], and much fewer than 1% are available in molecular identification reference libraries [63–65]. Thus, one cause of our current restricted size of small molecule identification libraries is due to the limited number of molecules available for purchase as authentic reference material. Even if all molecules were known and available for purchase, the time and cost to analyze these for building reference libraries would be prohibitive [66, 67]. The fields of metabolomics and exposomics, and small molecule identification generally, must overcome the significant, longstanding obstacle in the field: the absence of analytical methods for comprehensive and unambiguous identification of small molecules without reliance on reference data obtained from analysis of chemical standards [68–70].

For molecular properties that are consistently calculable with a known (preferably low) error, it is possible to create *in silico* reference libraries in order to reduce reliance on authentic chemical standards [70]. Several analytical methodologies, such as those based on chromatography coupled with MS [71–73] and NMR [74] have demonstrated feasibility for compound identification based on predicted properties. NMR's ability to be non-destructive and easily quantifiable makes it a unique tool for identifying novel compounds and handling complex metabolite mixtures without the need of chemical separation [75]. For example, MS/MS spectra yield reasonable accuracy for predictions of molecular properties and can be coupled with machine learning methods [76] but limited to short lists of small molecules [77, 78]. Quantum chemical applications such as infrared spectra [79], molecular collisional cross sections (CCS) [80, 81] and NMR chemical shifts [82–85], are promising for the calculations of molecular attributes. For example, coupling calculated mass and CCS has contributed to successful chemical identification of *cis/trans* isomers [86, 87], as well as isomers in complex synthetic samples [88]. For studies specifically using NMR, quantum chemical simulations for the prediction of spectra have been a valuable tool for the community. In the last two decades, density functional theory (DFT), an exceptionally well-established approach for high-throughput chemical calculations with the advantage of high performance for less computational cost, has been widely applied to predict NMR chemical shifts [89–91] of molecules and conformers [92–94] in different custom solvent conditions [95–97]. Furthermore, structural elucidation is one of the most practical uses of NMR, and it is common to utilize NMR chemical shift calculations along with experimental shifts to identify compound mixtures [98–100] and to aid

reassignment of structures or stereostructure assignment [101–103].

Currently, the use and acceptance of predicted NMR chemical shifts is limited due to an incomplete understanding of the required accuracy of such predictions for confident molecular identification. It has already been demonstrated that heuristic/empirical approaches for chemical shift predictions are generally of low accuracy compared to quantum chemical calculation-based methods (e.g. DFT) [83, 104, 105]. For DFT approaches, the factors that significantly affect the accuracy of predicted ^{13}C and/or ^1H NMR chemical shifts are the optimization level of molecular geometry [106–108], the use of different DFT theories [109–112], implicit and explicit solvation models [113–115], unique molecular properties of metabolites [116–121], etc. Agreement between predicted and experimental chemical shifts can be improved when (i) the basis set is enlarged [104, 122], (ii) the quality of the method is improved for geometry optimization [123, 124], (iii) a scaling procedure is employed [125, 126], (iv) conformational sampling is applied [127], and (v) solvation is taken into account appropriately [128, 129]. However, the question of what level of accuracy is required for calculated NMR chemical shifts when using these as reference spectra for molecular identification remains largely unexplored.

In this study, we investigate the accuracy and/or level of confidence in predicted NMR chemical shifts required to identify small molecules using reference libraries of varying size. Specifically, we present a detailed study on the role of accuracy in the prediction of ^{13}C and ^1H NMR spectra for confident metabolite identification in solution phase using a chloroform and water continuum model. We estimate the minimum and maximum error limits which hinder or enable ^{13}C and ^1H NMR chemical shift predictions to unambiguously identify molecular structures. In this study, we discuss two cases—simple and complex samples—using 11,716 small molecules taken from the HMDB [18]. We cover different chemical functional groups and explore the results to provide statistics for libraries of different sizes.

Materials and methods

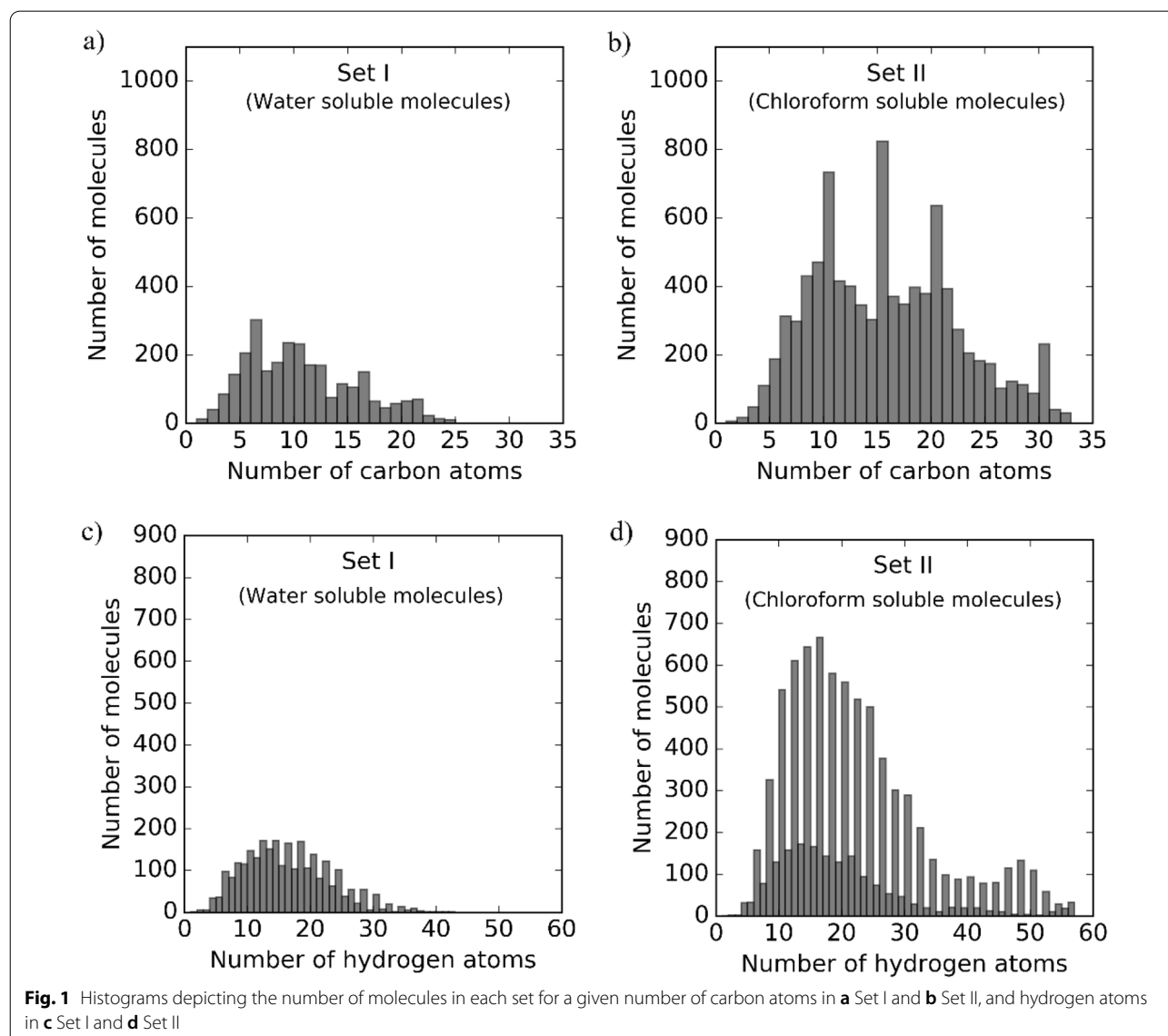
Molecule sets

Two sets of molecules taken from HMDB 4.0 [56] and distinguished by their reported partition coefficients were simulated, one in water (Set I) and a second in chloroform (Set II) as the solvent. The included compounds were not in salt forms, consist only of C, H, O, N, P and S atoms, and are in the molecular weight range of 27 to 500 Da. Set I, the water solvated set, contains 2,723 molecules (29,489 carbon and 45,426 hydrogen nuclei in total across all molecules) and spans a wide range of

structure-based chemical classes and chemical functionalities including organic acids, organonitrogen compounds, nucleosides, nucleotides, organoheterocyclic compounds, carboxylic acids, organooxygen compounds, and benzenoids as determined by the hierarchical chemical classification scheme, ClassyFire [130]. Set II, the chloroform solvated set, contains 8,990 molecules (138,535 carbon and 191,327 hydrogen nuclei in total across all molecules) and also spans a broad range of chemical functionalities including organic compounds, organic acids, lipids, benzenoids, and organoheterocyclic compounds. Figure 1 compares the number of molecules containing a given amount of carbons and hydrogens for Sets I and II. The molecules and their geometries in both Sets are provided in the Additional file 1.

Computational details

The NMR chemical shifts for all molecules in this study were calculated using the In Silico Chemical Library Engine (ISiCLE) [131] (see github.com/pnnl/isicle for the latest version of ISiCLE). ISiCLE is an automated pipeline for high-accuracy chemical property calculation, implemented using the Snakemake workflow management system [132]. This pipeline takes SMILES [133] (a line notation representation of molecule structure) as input, generates initial 3D molecular conformations, and subsequently optimizes this initial structure and calculates chemical properties through quantum chemistry via NWChem [134] (an open-source, high-performance computational chemistry software developed at PNNL). For this study, all molecules were initially optimized in solvent using the computationally inexpensive B3LYP [135, 136] with 3-21G basis set [137–139]. We chose this level of theory due to our available computational resources, particularly considering the treatments for the geometry optimization of over 11 k molecules. It is known that the 3-21G basis set for geometry optimization is not adequate to obtain high accuracy in NMR chemical shift calculations [127, 140, 141], but in this study it is only used to simulate NMR spectral data in order to obtain a reasonable representative distribution of (likely moderate accuracy) chemical shifts. Assessment of the best computational approaches to maximize accuracy of NMR chemical shift calculations is beyond the scope of this study. To test whether the NMR spectral data is statistically affected or not by using any other DFT method, the isotropic shielding values of 5 randomly chosen molecules in different shapes and sizes from Sets I and II were calculated using 3 different DFT methods. The shielding values were observed to be shifted in the same direction following the same pattern. Further details are given in Additional file 2. The inclusion of solvent is via the COnductor-like Screening MOdel (COSMO) [142] solvation modeling.



NMR isotropic shieldings were calculated for all optimized molecules at the B3LYP/cc-pVDZ [139, 143] level of theory. Based on our previous assessment [131], this method provides reliable chemical shifts [112] and yields isotropic shieldings with a reasonably low computational cost [144]. The gauge-invariant atomic orbital (GIAO) approach [145] was used to compute ^{13}C and ^1H NMR chemical shifts. The computed chemical shifts are provided in Additional file (available upon author request).

Algorithm

Various scoring approaches have been proposed for the analysis of chemical shifts and comparisons of DFT methods. The most common criteria in the literature quantifying the agreement between calculated and

experimental data are mean absolute error (MAE) (Eq. 1), root mean square error (RMSE) (Eq. 2), corrected mean absolute error (CMAE) (Eq. 3), and correlation coefficients (e.g., the Pearson correlation coefficient).

$$\text{MAE} = \frac{\sum_{i=1}^N |\delta_{\text{exp}} - \delta_{\text{calc}}|}{N} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (\delta_{\text{exp}} - \delta_{\text{calc}})^2}{N}} \quad (2)$$

$$\text{CMAE} = \frac{\sum_{i=1}^N |\delta_{\text{exp}} - (\delta_{\text{calc}} - b)/m|}{N} \quad (3)$$

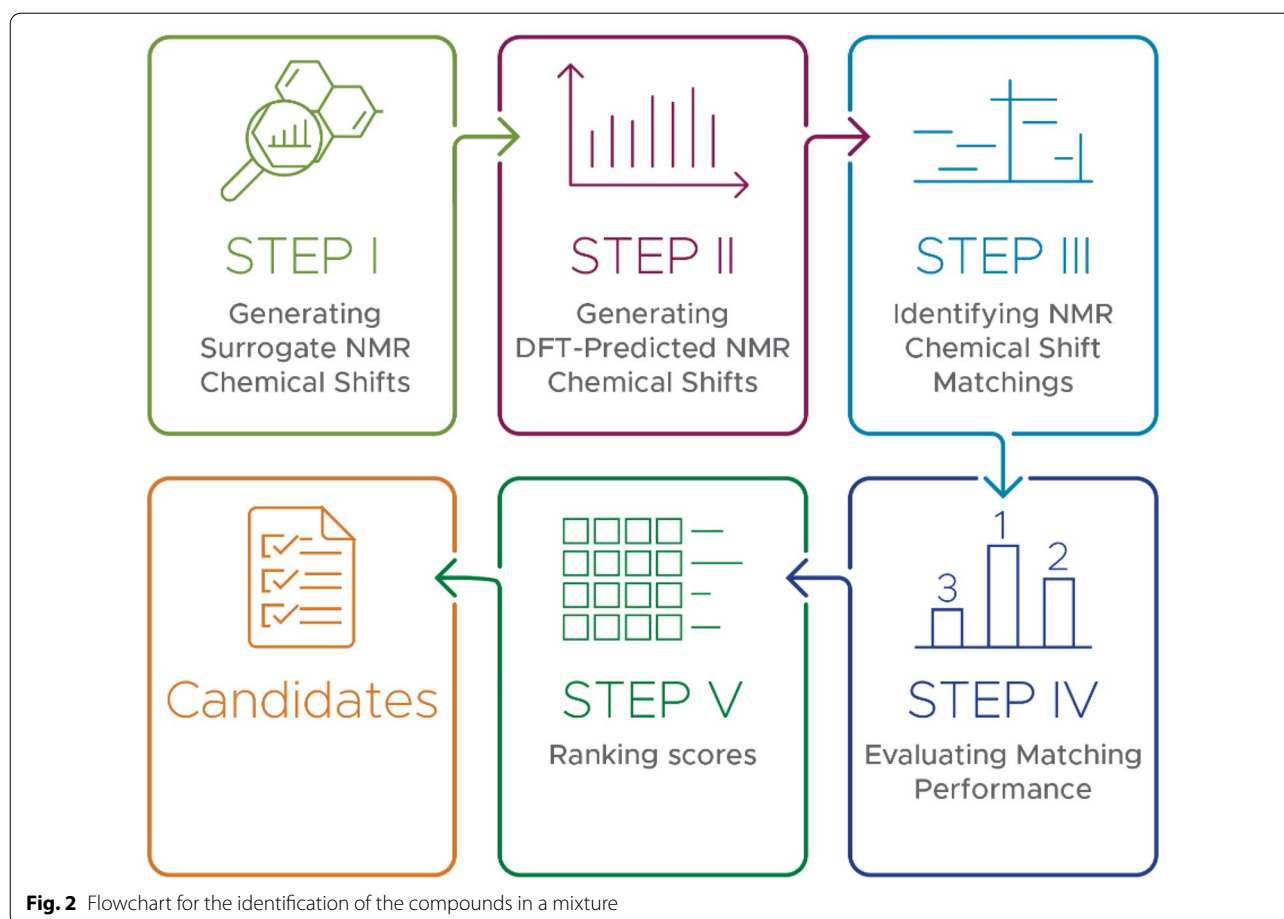


Fig. 2 Flowchart for the identification of the compounds in a mixture

where δ_{exp} is the experimental chemical shift, δ_{calc} is the calculated chemical shift, N is the number of nuclei, and m and b denote slope and intercept of the calculated shifts with respect to experimental shifts.

To identify the compounds in a mixture, our approach follows the steps in the flowchart presented in (Fig. 2). In Step I, NMR chemical shifts of all molecules are calculated as described above. Since we do not have experimental NMR data for the 11 thousand molecules in our two sets, in step II we create representative NMR data for comparisons: the calculated NMR spectra (generated in Step I) are considered as surrogate experimental shifts data and new lists of chemical shifts are created synthetically by adding Gaussian distributed noise. Although the error distributions of NMR chemical shifts were reported to also obey a student t-distribution in other studies [131, 146–149], we assume errors for both carbons and protons follow a Gaussian distribution [144, 150] with mean μ and standard deviation σ . Unless otherwise stated, the mean is assigned as 0, since the errors of scaled ^{13}C and ^1H NMR chemical shifts are equally likely to be positive or negative [144, 147]. In this study, σ is taken in the range of 0.5–50 ppm and 0.1–10 ppm for ^{13}C

and ^1H chemical shifts with increment of 0.05 ppm and 0.01 ppm, respectively. Simply, we assume that our initial (non-noise-added) calculated chemical shifts (“surrogate experimental data”) represent the distribution, but not necessarily the accuracy of authentic experimental chemical shifts, and that the addition of zero-mean Gaussian noise to create synthetic data with a defined error allows us to explore how the accuracy of real calculated chemical shifts can affect identification rates. This approach is similar to that taken in other successful studies [151–153]

In Step III, each molecule taken from the computed data is searched back against the surrogate experimental data. First, the experimental chemical shifts of an unknown molecule are matched to the computed chemical shifts of every single molecule to find the best match, based on minimizing the distance between two sets of chemical shifts. To do this, we used the Munkres assignment algorithm [154–156], which gives the minimum distance score (i.e. error) of two sets, within a feasible computational time bounded by a polynomial expression [157]. The Munkres algorithm minimizes the total error or summation of squared differences between each assignment. It is based on the following principle:

Let S_1 and S_2 be two separate lists of chemical shifts consisting of N and M elements, respectively. Let us construct an M -by- N matrix

$$\begin{bmatrix} (s_1 - b_1)^2 & \cdots & (s_M - b_1)^2 \\ \vdots & \ddots & \vdots \\ (s_1 - b_N)^2 & \cdots & (s_M - b_N)^2 \end{bmatrix}$$

where s_i is the m th element of S_1 , b_j is the n th element of S_2 and $M \leq N$. We have M elements to be assigned to N elements on a one-to-one basis where the assignments constitute an independent set of the M -by- N matrix. Then, the Munkres algorithm models an assignment problem, which returns the least-sum of elements of the matrix, choosing only one element from each row and column. In our case, this indicates the best possible matching, which will be used in the next step.

In Step IV, for each molecule, to determine which set of experimental data best matches to the computed one, the similarity of two sets of assigned chemical shifts is quantified by a distance score. There is no perfect score (i.e. zero error) between two sets (e.g., in practice, there is always some amount of error expected between experimental and computed shifts). A critical issue is finding a method to quantify the error such that it always yields the best match at the top when the list of scores are sorted from most to least likely. In addition to the most popular ways to express chemical shift errors (i.e. MAE and RMSE), we believe that an indication of how confident a matching set is can be expressed better in terms of RMSE and probability. Smith et al. performed a sophisticated systematic study for addressing the issue of the best parameter, and proposed DP4 [147], which is used when experimental NMR data is to be used to identify one molecule out of an arbitrarily large library of many possible structures. DP4 is based on conditional probability and/or Bayes' theorem—the key factor increasing the certainty of results. While we found DP4 to give slightly better rankings for pure samples than RMSE, we also found it to be computationally much more intense than RMSE. We also believe DP4 is not convenient for ranking matches in impure/complex samples. Therefore, we use RMSE in this study. Further details are given in Additional file 1.

Note that the RMSE ranges differ for carbon and proton. For the cases when carbon and proton are used together for identifying molecules, each RMSE is calculated separately and their geometric means are taken to get a single score for the molecule. The geometric mean is used to normalize the RMSEs, so the error associated with carbon does not dominate that of the proton for cases where both nuclei are used together.

Finally, in Step V, all resulting scores are sorted in ascending order, yielding a list of molecules starting from the most likely to the least likely to be found in the mixture. The ranks and scores of each molecule are reported. In this study, a rank of 1 (top of list) is synonymous with positive molecule identification.

For this study, we considered the case when (1) proton chemical shifts are used alone for identification, (2) carbon chemical shifts are used alone, and finally, (3) when both nuclei are used together.

The automated workflow and all scripts, written in Python, are provided in Additional file 4.

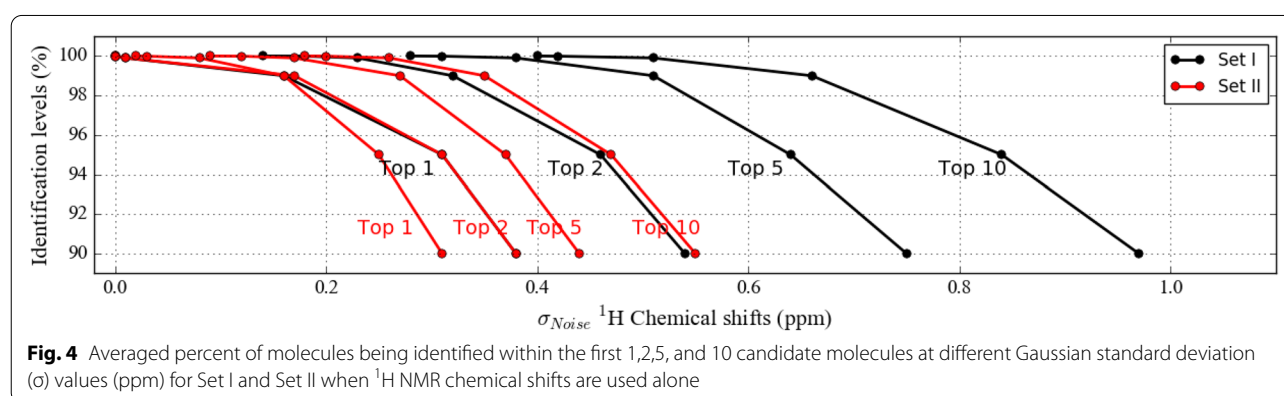
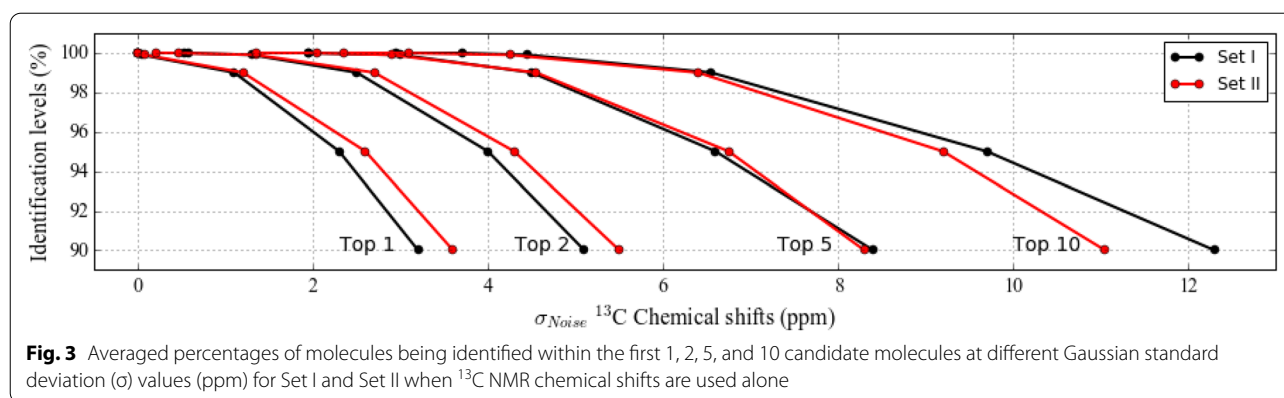
Results and discussion

Robust and comprehensive metabolite identification using calculated NMR chemical shifts requires assessments of the accuracies of the *in silico* approaches used and that must have validated error ranges. We investigated the level of accuracy required to identify small molecules in NMR libraries. We performed a comprehensive analysis on the extent of accuracy in the predicted ^{13}C and ^1H NMR chemical shifts using 11,716 small molecules taken from the HMDB. We analyzed the limits (upper and lower) of error for confident metabolite identification. in two solution phases: chloroform and water. We discussed the possible error ranges in predicted NMR chemical shifts allowing to achieve reasonably confident identification in 2 types of samples: (i) pure uniform sample, and (ii) complex sample. We performed our runs for 190 different error ranges (i.e. σ , Gaussian standard deviation) and repeated the experiments 16 times for each case. Unless otherwise stated, all analyses were performed for each molecule in the two sets. We report the average results for i) ^{13}C chemical shifts alone, ii) ^1H chemical shifts alone, and iii) ^{13}C and ^1H chemical shifts used together for identification. We report the average percentage of molecules successfully identified (i.e. rank is 1) for Set I (water soluble molecules) and Set II (chloroform soluble molecules).

Case I: Pure sample

In this case, let us assume we have a spectrum from a single compound and an array of carbon (^{13}C) and/or proton (^1H) NMR chemical shifts. This case involves selecting only the molecules having the exact number of carbon and/or proton chemical shifts from the database to match the experimental spectrum. This narrows the list of candidate molecules.

Figures 3 and 4 show the identification results of 90% to 100% of the molecules of both sets in the top 10 hits (Top 10) for carbons and protons used independently.



As an example, for identifying 90% of the molecules in the first hit (Top 1), ^{13}C chemical shift errors should be below 3.2 and 3.6 ppm for Set I and Set II, respectively. Likewise, for ^1H chemical shifts, when the MAE is at most 0.38 ppm for the both sets, there is a 90% chance that the correct identification will be made as the first hit. It is possible to correctly identify 99% of the molecules when the noise is at most 1.1–1.2 and 0.16–0.17 ppm for ^{13}C and ^1H chemical shifts, respectively. The molecule of interest has a chance to be among the first two candidate matches (Top 2) when ^{13}C and ^1H chemical shift errors are 0.53 ppm and 0.21 ppm, and 0.14 ppm and 0.02 ppm for Set I and Set II, respectively. However, for these sets of molecules, 100% of identification is not possible when ^{13}C and ^1H chemical shifts are used alone. The higher quality versions of Figs. 3 and 4, and the full list including 50–100% of identification is given in the Additional file (available upon author request).

Figure 5 shows where the molecules rank in identification lists for a comprehensive identification analysis for Set I and Set II, plotted against carbon and proton errors when ^{13}C and ^1H data are used together for identification. The plots show how the probability of a molecule being correctly identified changes with chemical shift errors.

The contour lines represent different levels of identification with respect to carbon (y-axis) and proton (x-axis) errors. The color bars show the ranking distributions along the ranges of carbon (0–50 ppm) and proton errors (0–10 ppm). The contour lines are represented in a reciprocal relationship (Eq. 1) (see Additional file 3 for further information). Therefore, on each contour line, it is possible to have a list of combinations for a range of carbon and proton errors. For example, for 90% of identification, the carbon and proton errors (ppm) could be (3 and 10), (5 and 0.92), or (6 and 0.7), respectively, out of many combinations. This reciprocal relationship also gives a trade-off between the carbon and proton errors such that it is possible to skip expensive ^{13}C chemical shifts over highly accurate ^1H chemical shifts, and vice versa.

$$\text{MAE}(^{13}\text{C}) = a/(\text{MAE}(^1\text{H}) - b) + c \quad (4)$$

Each contour line has an optimum point which represents a trade-off point (reported in Table 1). At these points on the curves, the cumulative errors of carbon and proton are minimum (note that ranges for carbon and proton errors are normalized). A fascinating but

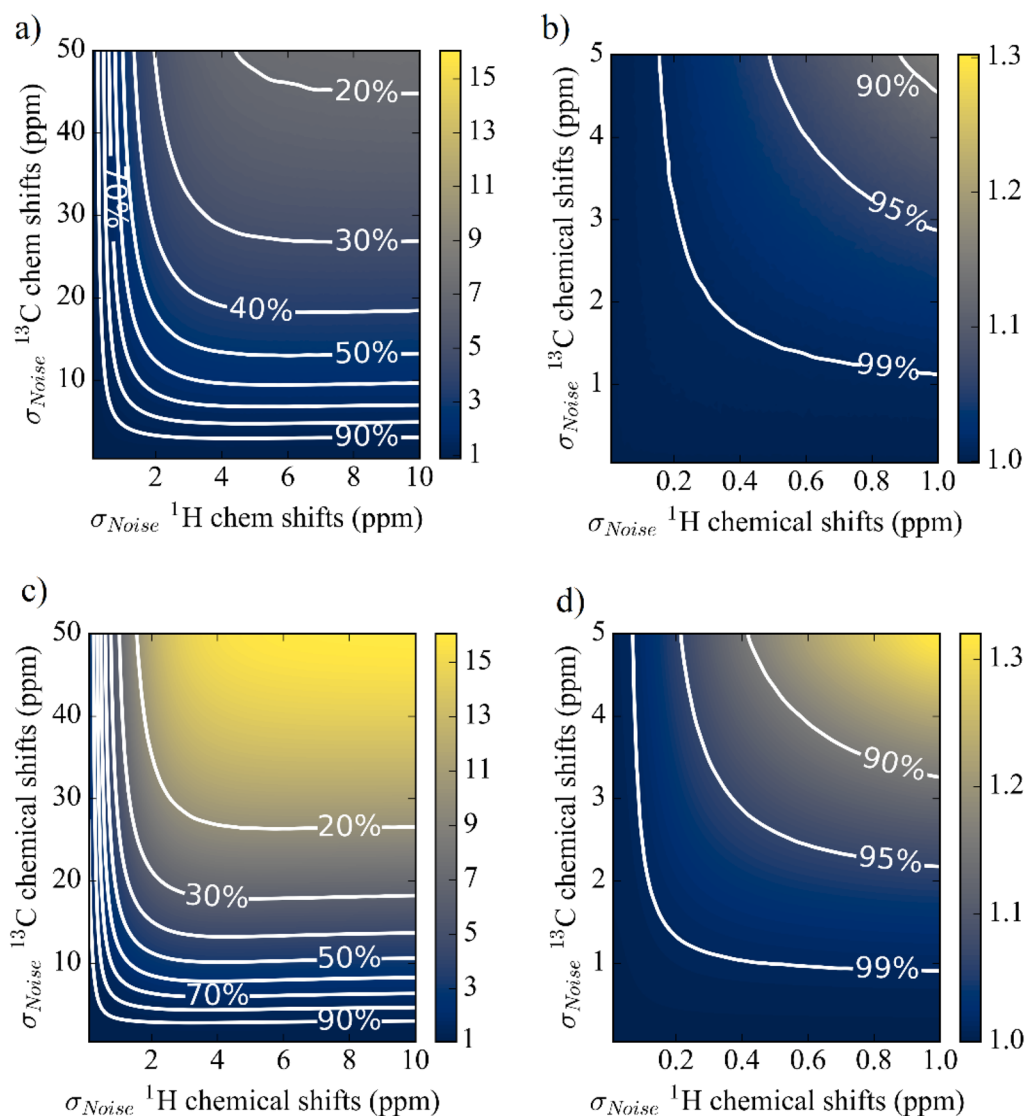


Fig. 5 Mean of ranks with respect to the carbon and proton errors and contour lines for the different level of identification ratios when carbons and protons are used together for **a, b** Set I (water soluble molecules) and **c, d** Set II (chloroform soluble molecules). **b** and **d** are the zoomed versions of **a** and **c**, respectively. The color bars represent the rankings

Table 1 Optimum trade-off MAEs at different Gaussian standard deviation (σ) values (ppm) for Set I and Set II when ^{13}C and ^1H NMR chemical shifts are used together for identification

Percentile (%)	Set I (Water soluble molecules) σ (ppm)		Set II (Chloroform soluble molecules) σ (ppm)	
	^{13}C	^1H	^{13}C	^1H
99	2.02	0.30	1.64	0.43
95	4.21	0.57	4.44	0.53
90	6.16	0.72	5.82	0.70

not unexpected observation here is that the chances of molecules being successfully identified are doubled when ^{13}C and ^1H chemical shifts are used together. Thus, compared to the previous case when ^{13}C and ^1H chemical shifts are used independently, using more information increases the chance of successful identification. The full list of trade-off points including 50–99% is reported in the Additional file (available upon author request).

It is observed the ranks range from 1 to 7 for Set I and 1 to 16 for Set II. The difference in ranges source from different sized molecule sets and differences in

standard deviation and variance of ^{13}C and ^1H chemical shifts. The standard deviations of ranks are shown in the Additional file 2.

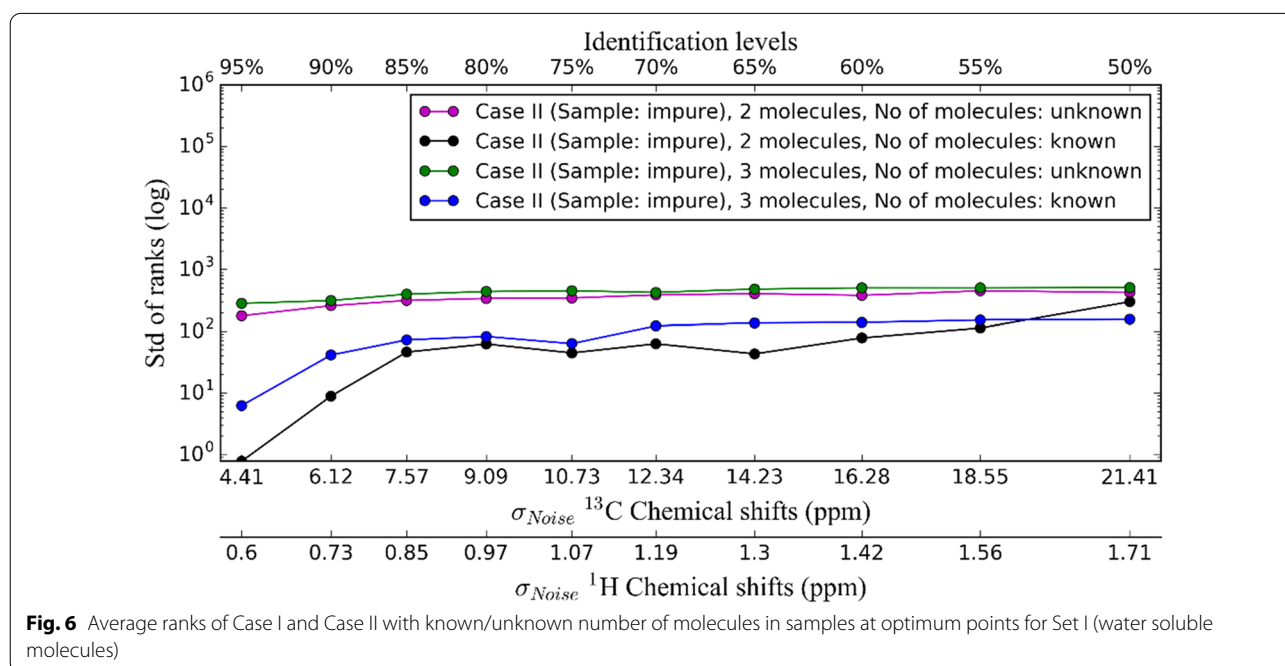
Case II: Impure sample

A continuing grand challenge for NMR-based metabolomics is dealing with the spectral complexity in analysis of mixtures. An NMR spectra can have a combination of thousands of distinct resonances belonging either to the main compound or to impurities. Here, we used an approach very similar to a quantitative metabolomics approach in which identification and quantification are based on the underlying assumption that any given sample spectrum is the sum of individual spectra of pure metabolites found in the mixture. The spectrum of interest is compared to a library of pure compound spectra by properly matching and fitting the reference peaks. The reference libraries need to be prepared from NMR spectra of pure metabolites at a precisely known and controlled pH and temperature. Especially, peaks of water or some endogenous metabolite are pH, temperature and salt-sensitive, which frequently leads to errors. In this study, we disregarded the effects of pH and temperatures, and distortions, artifacts and noise in signals. We performed our analysis based on the assumption that the spectrum of every single compound in the mixture is a sub-spectrum stored in the reference database.

Let us assume we have an impure sample consisting of unknown number of compounds and carbon and/or proton NMR chemical shift data for the sample. In contrast

to Case I, here we consider an n-tuple of molecules to be the list of candidates in the sample consisting of n number of molecules. Unlike Case I, the sequence of chemical shifts to be matched in the reference library do not necessarily have the same size of candidates; instead any molecule having equal or less ^{13}C and ^1H NMR chemical shifts in the reference library has a chance to be a candidate. For instance, if we have a sample of 2 molecules with c_1 and c_2 number of carbons and h_1 and h_2 number of protons, respectively, only the pairs having a sum of $c_1 + c_2$ carbons and $h_1 + h_2$ protons are the candidates and the chemical shifts of an atom can only belong to one of two candidates.

Compared to Case I, not only does the list of candidate molecules expand but matching two sets of data of different size is also not straightforward, making it even more challenging. Because of this, we did not examine this case for different Gaussian noise levels in detail as we did in Case I. We performed our runs for mixtures of 2 and 3 compounds. We report the results of this case only for a specific set of Gaussian noises (the optimum trade-off MAEs of ^{13}C and ^1H NMR chemical shifts reported for Set I in Table 1). Unless otherwise specified, we refer the mixtures of 2 and 3 compounds as pairs and triplets, respectively. In Fig. 6, the averaged ranks are shown for molecule pairs and triplets for all the optimum MAEs. Compared to the case of pure samples (Case I), the probability of identification decreases from 95 to 0% (pairs) and 6% (triplets) when the ^{13}C and ^1H NMR chemical shift errors



are 4.41 ppm and 0.6 ppm, respectively. So, the identification chance is quite low (green and purple lines in Fig. 6) even when the ^{13}C and ^1H NMR chemical shift errors are low. We then investigated what happens if the number of compounds in the sample is known. At first this seems counter intuitive, but the probability of identification is increased to 84% (from 0%—pairs) and 68% (from 6%—triplets) when the ^{13}C and ^1H NMR chemical shift errors are 4.41 ppm and 0.6 ppm, respectively. The average identification chances increase by 83% and 91% (blue and black lines in Fig. 6). Determining the number of compounds in a sample may be possible using additional orthogonal data. For example, multidimensional NMR experiments or MS may aid in determining the number of high concentration molecular candidates in a sample. Integrating NMR and MS can provide improved identification and quantification of a larger number of metabolites, as in Case II. [158, 159]. This is still, however, less than Case I by 84% and 93% for pairs and triplets, respectively (red line in Fig. 6). Standard deviations of ranks and computational times of runs are given in the Additional file 1.

Case II was performed only for the smaller molecule set, Set I (water soluble molecules), and not for the larger set, Set II (chloroform soluble molecules), due to the high computational time demands.

NMR spectroscopy is one of the main methods used for identifying the structure of metabolites. Besides the usual parameters (i.e. ^{13}C and ^1H NMR chemical shifts), other major NMR parameters (i.e. spin–spin coupling constants and ^{15}N , ^{17}O , and other nuclei chemical shifts) can alternatively be used for structure identification. We believe the use of any other property will significantly improve molecular identification. In this initial study, we did not test the effect of using additional information that can be collected using NMR (e.g., J-couplings and peak shape). However, most currently available databases provide only ^{13}C and ^1H NMR chemical shifts, and J-couplings, multi-dimensional spectra, etc. are missing for many molecules. There is rapid progress in the use of 2D NMR models (i.e. COSY, HSQC, and HMBC) which aids interpretation of spectrum and leads to less ambiguity in the spectral assignments and allows more reliable identification. 2D NMR techniques are proven to overcome the problem of insufficient spectral resolution and spectral redundancy. 2D NMR experiments provide additional information (i.e. couplings between magnetic nuclei) and solve the problem of overlapping peaks. Thus, it allows identification of metabolites that otherwise remain undetected. Multi-dimensional spectra prediction can be obtained using spin dynamics simulation libraries

(i.e. SPINACH [160]) coupled with DFT calculations. We are currently assessing the present limits of such automated workflows for accelerating confident, accurate, and fast metabolite identification.

Conclusion

Global comprehensive compound identification in complex samples will revolutionize understanding of the role of important compounds in chemical, environmental and biological studies. A major limitation is that the vast majority of metabolites are not available in current identification libraries, nor available for purchase as authentic reference material. It is not economically and practically feasible to identify hundreds of thousands of metabolites in laboratories to establish small molecule reference libraries. To address this, *in silico* small molecule libraries are currently the only reasonable solution to move toward comprehensive identification of all molecules in complex samples.

We performed an extensive statistical analysis on the effect of ^{13}C and ^1H NMR chemical shift calculation errors, in water and chloroform solvents, on the ability to make correct identification from *in silico* libraries. For pure samples, the required accuracy levels are feasible, promising the establishment of large scale metabolomic NMR *in silico* libraries. 90% or more of these molecules in a pure sample can be successfully identified when errors of ^{13}C and ^1H NMR chemical shifts are below 6 ppm and 0.5 ppm, respectively. This shows great potential of future use and reliability of predicted NMR chemical shifts in molecule identification for pure samples.

Compared to pure sample identification, it may require complementary information for complex samples in order to correctly identify constituent compounds. The water-soluble molecules in a complex sample have a chance of 68% and 84% (it is 95% for pure samples) to be identified for pairs and triplets, respectively when errors of ^{13}C and ^1H NMR chemical shifts are below 4.41 ppm and 0.6 ppm. The possibility of identification increases by 90% when the number of molecules are known beforehand, corroborating other findings that significant potential for parallel MS analysis [161]. This increased confidence in our results indicates the value of adding multiple molecular or chemical properties and using additional measured or accurately predicted information for comprehensive identification of metabolites.

This study provides valuable insight into the practicality and applicability of potential *in silico* small molecule NMR databases. The rapid innovations in metabolite identifications through the recent advances in computation and data integration in both NMR and MS/NMR analytical and computational methods will aid the full

metabolome composition assignment in complex sample identification.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13321-022-00587-7>.

Additional file 1. Set I - Water Soluble Molecules.

Additional file 2. Set II - Chloroform Soluble Molecules.

Additional file 3. Python Scripts.

Additional file 4. Supplementary Information Document.

Acknowledgements

Not applicable.

Authors' contributions

YY performed all calculations, analyzed the results, created the figures, and was the primary author. NG assisted with the DFT calculations. TOM and RSR oversaw the research, helped with analysis, and provided the funding for the study. All authors contributed to writing.

Funding

This work was supported by the Microbiomes in Transition (MinT) Initiative as part of the Laboratory Directed Research and Development Program at PNNL. Additional support was provided by the National Institutes of Health, National Institute of Environmental Health Sciences Grant no. U2CES030170. PNNL is a multi-program national laboratory operated by Battelle for the DOE under contract DE-AC05-76RLO 1830.

Availability of data and materials

Molecule MOL files and DFT output files are included in the Additional files, along with Python processing code. Any other data is freely available upon request. The Additional files are available from the authors, upon request.

Declarations

Competing interests

The authors declare no competing interests.

Author details

¹The Gene and Linda Voiland School of Chemical Engineering and Bioengineering, Washington State University, Pullman, WA, USA. ²Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA.

Received: 3 December 2020 Accepted: 6 February 2022

Published online: 22 September 2022

References

- German JB, Hammock BD, Watkins SM (2005) Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics* 1(1):3–9
- Wishart DS (2007) Current progress in computational metabolomics. *Brief Bioinform* 8(5):279–293
- Shulaev V (2006) Metabolomics technology and bioinformatics. *Brief Bioinform* 7(2):128–139
- Kosmidis AK et al (2013) Metabolomic fingerprinting: challenges and opportunities. *Crit Rev Biomed Eng* 41(3):205–221
- Nicholson JK, Wilson ID (2003) Opinion: understanding “global” systems biology: metabolomics and the continuum of metabolism. *Nat Rev Drug Discov* 2(8):668–676
- Winnike JH et al (2010) Use of pharmaco-metabolomics for early prediction of acetaminophen-induced hepatotoxicity in humans. *Clin Pharmacol Ther* 88(1):45–51
- Holmes E et al (2008) Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* 453(7193):396–400
- Beckonert O et al (2007) Metabolic profiling, metabolomic and metabolomics procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* 2(11):2692–2703
- Nicholson JK, Lindon JC, Holmes E (1999) “Metabonomics”: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29(11):1181–1189
- Nicholson JK et al (1995) 750 MHz ¹H and ¹H–¹³C NMR spectroscopy of human blood plasma. *Anal Chem* 67(5):793–811
- Smith CA et al (2006) XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. *Anal Chem* 78(3):779–787
- Dettmer K, Aronov PA, Hammock BD (2007) Mass spectrometry-based metabolomics. *Mass Spectrom Rev* 26(1):51–78
- Want EJ, Cravatt BF, Siuzdak G (2005) The expanding role of mass spectrometry in metabolite profiling and characterization. *ChemBioChem* 6(11):1941–1951
- Dunn WB, Bailey NJ, Johnson HE (2005) Measuring the metabolome: current analytical technologies. *Analyst* 130(5):606–625
- Hollywood K, Brison DR, Goodacre R (2006) Metabolomics: Current technologies and future trends. *Proteomics* 6(17):4716–4723
- Moco S et al (2007) Metabolomics technologies and metabolite identification. *Trac-Trends Anal Chem* 26(9):855–866
- Smith CA et al (2005) METLIN: a metabolite mass spectral database. *Ther Drug Monit* 27(6):747–751
- Wishart DS et al (2013) HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Res* 41(Database issue):D801–D807
- Ulrich EL et al (2008) BioMagResBank. *Nucleic Acids Res* 36(Database):D402–D408
- Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87(11):1123–1124
- Tautenhahn R et al (2012) XCMS online: a web-based platform to process untargeted metabolomic data. *Anal Chem* 84(11):5035–5039
- Williams AJ (2008) A perspective of publicly accessible/open-access chemistry databases. *Drug Discov Today* 13(11–12):495–501
- Sitzmann M, Filippov IV, Nicklaus MC (2008) Internet resources integrating many small-molecule databases. *SAR QSAR Environ Res* 19(1–2):1–9
- Kutzler FW et al (1983) Charge-Density and bonding in (5,10,15,20-tetramethylporphyrinato)nickel(II)—a combined experimental and theoretical study. *J Am Chem Soc* 105(10):2996–3004
- Stimpson DI, Cann JR (1981) A combined theoretical and experimental study of the interaction of metrizamide with proteins. *Arch Biochem Biophys* 211(1):403–412
- Cripps SC, Orton RS, Carroll JE (1974) Combined theoretical and experimental studies of a push-pull trapatt circuit. *Int J Electron* 37(1):1–21
- Gaulton A et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40(Database issue):D1100–D1107
- Izgi T et al (2007) FT-IR and NMR investigation of 2-(1-cyclohexenyl) ethylamine: a combined experimental and theoretical study. *Spectrochimica Acta Part A Mol Biomol Spectrosc* 68(1):55–62
- de Matos P et al (2010) Chemical entities of biological interest: an update. *Nucleic Acids Res* 38:D249–D254
- Kwan EE, Liu RY (2015) Enhancing NMR prediction for organic compounds using molecular dynamics. *J Chem Theory Comput* 11(11):5083–5089
- Knox C et al (2011) DrugBank 3.0: a comprehensive resource for “Omics” research on drugs. *Nucleic Acids Res* 39:D1035–D1041
- Ulrich EL et al (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408
- Wishart DS et al (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* 37(Database issue):D603–D610
- Blum LC, Reymond JL (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J Am Chem Soc* 131(25):8732–8733
- Jewison T et al (2014) SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Res* 42(Database issue):D478–D484
- Frolkis A et al (2010) SMPDB: the small molecule pathway database. *Nucleic Acids Res* 38(Database issue):D480–D487

37. Richard AM, Williams CR (2002) Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat Res* 499(1):27–52
38. Guo AC et al (2013) ECMDDB: the *E. coli* metabolome database. *Nucleic Acids Res* 41(Database issue):D625–D630
39. Sajed T et al (2016) ECMDDB 2.0: a richer resource for understanding the biochemistry of *E. coli*. *Nucleic Acids Res* 44(D1):D495–501
40. Keseler IM et al (2017) The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res* 45(D1):D543–D550
41. Scalbert A et al (2011) Databases on food phytochemicals and their health-promoting effects. *J Agric Food Chem* 59(9):4331–4348
42. Fahy E et al (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res* 50:59–S14
43. Caspi R et al (2018) The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 46(D1):D633–D639
44. MolMall. [cited 2019 8/1]; <http://www.molmall.net/>.
45. Banerjee P et al (2015) Super Natural II-a database of natural products. *Nucleic Acids Res* 43(D1):D935–D939
46. Wishart D et al (2015) T3DB: the toxic exposome database. *Nucleic Acids Res* 43(Database issue):D928–D934
47. Lim E et al (2010) T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res* 38:D781–D786
48. Richard AM et al (2016) ToxCast chemical landscape: paving the road to 21st century toxicology. *Chem Res Toxicol* 29(8):1225–1251
49. Gu JY et al (2013) Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS ONE* 8(4):e62839
50. Sterling T, Irwin JJ (2015) ZINC 15-ligand discovery for everyone. *J Chem Inf Model* 55(11):2324–2337
51. Wishart DS (2011) Advances in metabolite identification. *Bioanalysis* 3(15):1769–1782
52. Xiao JF, Zhou B, Resson HW (2012) Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trac-Trends Anal Chem* 32:1–14
53. NIST 17 MS/MS Library. [cited 2019 05.01]. <https://www.sisweb.com/software/nist-msms.htm>.
54. The NIST 17 Mass Spectral Library. June 2017 [cited 2019 05.01]. <https://www.sisweb.com/software/ms/nist.htm#stats>.
55. The Human Metabolome Library (HML). [cited 2019 05.01]. <http://www.hmdb.ca/hml>.
56. Wishart DS et al (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 46(D1):D608–D617
57. Wishart DS et al (2007) HMDB: the human metabolome database. *Nucleic Acids Res* 35(Database issue):D521–D526
58. ZINC 15, a free database of commercially-available compounds. [cited 2019 05.01]. <http://zinc15.docking.org/>.
59. Sterling T, Irwin JJ (2015) ZINC 15-ligand discovery for everyone. *J Chem Inf Model* 55(11):2324–2337
60. Styczinski MP et al (2007) Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Anal Chem* 79(3):966–973
61. Staniek A, Woerdenbag HJ, Kayser O (2008) Endophytes: exploiting biodiversity for the improvement of natural product-based drug discovery. *J Plant Interact* 3(2):75–93
62. Tulp M, Bohlin L (2002) Functional versus chemical diversity: is biodiversity important for drug discovery? *Trends Pharmacol Sci* 23(5):225–231
63. Sumner LW et al (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics* 3(3):211–221
64. DeHaven CD et al (2010) Organization of GC/MS and LC/MS metabolomics data into chemical libraries. *J Cheminformatics* 2:1–12
65. Dobson CM (2004) Chemical space and biology. *Nature* 432(7019):824–828
66. Patti GJ et al (2013) A view from above: cloud plots to visualize global metabolomic data. *Anal Chem* 85(2):798–804
67. Weckwerth W (2003) Metabolomics in systems biology. *Annu Rev Plant Biol* 54:669–689
68. Salek RM et al (2013) The role of reporting standards for metabolite annotation and identification in metabolomic studies. *Gigascience* 2:2047–2217
69. Fiehn O et al (2007) The metabolomics standards initiative (MSI). *Metabolomics* 3(3):175–178
70. Beisken S, Eiden M, Salek RM (2015) Getting the right answers: understanding metabolomics challenges. *Expert Rev Mol Diagn* 15(1):97–109
71. Di Stefano V et al (2012) Applications of liquid chromatography-mass spectrometry for food analysis. *J Chromatogr A* 1259:74–85
72. Garcia A, Barbas C (2011) Gas chromatography-mass spectrometry (GC-MS)-based metabolomics. *Methods Mol Biol* 708:191–204
73. Schymanski EL et al (2014) Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* 48(4):2097–2098
74. Tang HR et al (2004) Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma. *Anal Biochem* 325(2):260–272
75. Nicholson JK, Wilson ID (2003) Understanding “global” systems biology: metabolomics and the continuum of metabolism. *Nat Rev Drug Discovery* 2(8):668–676
76. Kangas LJ et al (2012) In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics* 28(13):1705–1713
77. Allen F, Greiner R, Wishart D (2015) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* 11(1):98–110
78. Wolf S et al (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* 11:1–12
79. Bouteiller Y et al (2008) Transferable specific scaling factors for interpretation of infrared spectra of biomolecules from density functional theory. *J Phys Chem A* 112(46):11656–11660
80. Colby SM et al (2019) ISICLE: a quantum chemistry pipeline for establishing in silico collision cross section libraries. *Anal Chem* 91(7):4346–4356
81. Nuñez JR, et al (2018) Advancing Standards-Free Methods for the Identification of Small Molecules in Complex Samples. arXiv preprint arXiv:1810.07367.
82. Casabianca LB, De Dios AC (2008) Ab initio calculations of NMR chemical shifts. *J Chem Phys* 128(5):052201
83. Lodewyk MW, Siebert MR, Tantillo DJ (2012) Computational prediction of ¹H and ¹³C chemical shifts: a useful tool for natural product, mechanistic, and synthetic organic chemistry. *Chem Rev* 112(3):1839–1862
84. Hill DE, Vasdev N, Holland JP (2015) Evaluating the accuracy of density functional theory for calculating ¹H- and ¹³C-NMR chemical shifts in drug molecules. *Comput Theor Chem* 1051:161–172
85. Lomas JS (2016) ¹H-1 NMR spectra of alcohols in hydrogen bonding solvents: DFT/GIAO calculations of chemical shifts. *Magn Reson Chem* 54(1):28–38
86. Zheng XY et al (2017) Structural elucidation of cis/trans dicaffeoylquinic acid photoisomerization using ion mobility spectrometry-mass spectrometry. *J Phys Chem Lett* 8(7):1381–1388
87. Zheng XY et al (2017) Enhancing glycan isomer separations with metal ions and positive and negative polarity ion mobility spectrometry-mass spectrometry analyses. *Anal Bioanal Chem* 409(2):467–476
88. Nunez JR et al (2019) Evaluation of in silico multi-feature libraries for providing evidence for the presence of small molecules in synthetic blinded samples. *J Chem Inf Model* 59(9):4052–4060
89. Forsyth DA, Sebag AB (1997) Computed C-13 NMR chemical shifts via empirically scaled GIAO shieldings and molecular mechanics geometries. Conformation and configuration from C-13 shifts. *J Am Chem Soc* 119(40):9483–9494
90. Auer AA, Gauss J, Stanton JF (2003) Quantitative prediction of gas-phase C-13 nuclear magnetic shielding constants. *J Chem Phys* 118(23):10407–10417
91. Mothana B, Ban FQ, Boyd RJ (2005) Validation of a computational scheme to study ¹⁵N- and ¹³C-13 nuclear shielding constants. *Chem Phys Lett* 401(1–3):7–12
92. Saito H (1986) Conformation-dependent C-13 chemical-shifts—a new means of conformational characterization as obtained by high-resolution solid-state C-13 Nmr. *Magn Reson Chem* 24(10):835–852
93. Jaime C et al (1991) C-13 Nmr chemical-shifts—a single rule to determine the conformation of Calix[4]Arenes. *J Org Chem* 56(10):3372–3376
94. Yannoni CS et al (1991) C-13 Nmr-study of the C60 cluster in the solid-state—molecular-motion and carbon chemical-shift anisotropy. *J Phys Chem* 95(1):9–10

95. Malkin VG et al (1996) Solvent effect on the NMR chemical shieldings in water calculated by a combination of molecular dynamics and density functional theory. *Chem Eur J* 2(4):452–457
96. Casanovas J et al (2001) Calculated and experimental NMR chemical shifts of p-menthane-3,9-diols. A combination of molecular dynamics and quantum mechanics to determine the structure and the solvent effects. *J Org Chem* 66(11):3775–3782
97. Benzi C et al (2004) Reliable NMR chemical shifts for molecules in solution by methods rooted in density functional theory. *Magn Reson Chem* 42:S57–S67
98. Kiamco MM et al (2018) Structural and metabolic responses of *Staphylococcus aureus* biofilms to hyperosmotic and antibiotic stress. *Biotechnol Bioeng* 115(6):1594–1603
99. Dreyer DR et al (2012) Elucidating the structure of poly(dopamine). *Langmuir* 28(15):6428–6435
100. Xin DY et al (2017) Development of a C-13 NMR chemical shift prediction procedure using B3LYP/cc-pVDZ and empirically derived systematic error correction terms: a computational small molecule structure elucidation method. *J Org Chem* 82(10):5135–5145
101. Garcellano RC et al (2018) Isolation of tryptanthrin and reassessment of evidence for its isobaric isostere wrightiadione in plants of the wrightia genus. *J Nat Prod* 82(3):440–448
102. Kutateladze AG, Reddy DS (2017) High-throughput in silico structure validation and revision of halogenated natural products is enabled by parametric corrections to DFT-computed ¹³C NMR chemical shifts and spin-spin coupling constants. *J Org Chem* 82(7):3368–3381
103. Kutateladze AG, Krenske EH, Williams CM (2019) Reassignments and corroborations of oxo-bridged natural products directed by OSE and DU8+ NMR computation. *Angew Chem Int Ed Engl* 58(21):7107–7112
104. Jain R, Bally T, Rablen PR (2009) Calculating accurate proton chemical shifts of organic molecules with density functional methods and modest basis sets. *J Org Chem* 74(11):4017–4023
105. Perez M et al (2006) Accuracy vs time dilemma on the prediction of NMR chemical shifts: a case study (chloropyrimidines). *J Org Chem* 71(8):3103–3110
106. Barone G et al (2002) Determination of the relative stereochemistry of flexible organic compounds by ab initio methods: conformational analysis and Boltzmann-averaged GIAO C-13 NMR chemical shifts. *Chem Eur J* 8(14):3240–3245
107. Barone G et al (2002) Structure validation of natural products by quantum-mechanical GIAO calculations of C-13 NMR chemical shifts. *Chem Eur J* 8(14):3233–3239
108. Remya K, Suresh CH (2013) Which density functional is close to CCSD accuracy to describe geometry and interaction energy of small non-covalent dimers? A benchmark study using gaussian09. *J Comput Chem* 34(15):1341–1353
109. Zhao Y, Truhlar DG (2008) Improved description of nuclear magnetic resonance chemical shielding constants using the M06-L meta-generalized-gradient-approximation density functional. *J Phys Chem A* 112(30):6794–6799
110. Magyarfalvi G, Pulay P (2003) Assessment of density functional methods for nuclear magnetic resonance shielding calculations. *J Chem Phys* 119(3):1350–1357
111. Cimino P et al (2004) Comparison of different theory models and basis sets in the calculation of C-13 NMR chemical shifts of natural products. *Magn Reson Chem* 42:S26–S33
112. Tormena CF, da Silva GVJ (2004) Chemical shifts calculations on aromatic systems: a comparison of models and basis sets. *Chem Phys Lett* 398(4–6):466–470
113. Cramer CJ, Truhlar DG (1999) Implicit solvation models: equilibria, structure, spectra, and dynamics. *Chem Rev* 99(8):2161–2200
114. Wiitala KW, Hoyer TR, Cramer CJ (2006) Hybrid density functional methods empirically optimized for the computation of C-13 and H-1 chemical shifts in chloroform solution. *J Chem Theory Comput* 2(4):1085–1092
115. Reddy G, Yethiraj A (2006) Implicit and explicit solvent models for the simulation of dilute polymer solutions. *Macromolecules* 39(24):8536–8542
116. Smirnov SN et al (1996) Hydrogen deuterium isotope effects on the NMR chemical shifts and geometries of intermolecular low-barrier hydrogen-bonded complexes. *J Am Chem Soc* 118(17):4094–4101
117. Benedict H et al (1996) Hydrogen/deuterium isotope effects on the N-15 NMR chemical shifts and geometries of low-barrier hydrogen bonds in the solid state. *J Mol Struct* 378(1):11–16
118. Gidley MJ, Bociek SM (1988) C-13 Cp/Mas Nmr-studies of amylose inclusion complexes, cyclodextrins, and the amorphous phase of starch granules—relationships between glycosidic linkage conformation and solid-state C-13 chemical-shifts. *J Am Chem Soc* 110(12):3820–3829
119. Buckingham AD (1960) Chemical shifts in the nuclear magnetic resonance spectra of molecules containing polar groups. *Can J Chem Revue Canadienne De Chimie* 38(2):300–307
120. Osmialowski B, Kolehmainen E, Gawinecki R (2001) GIAO/DFT calculated chemical shifts of tautomeric species 2-Phenaclypyridines and (Z)-2-(2-hydroxy-2-phenylvinyl)pyridines. *Magnet Reson Chem* 39(6):334–340
121. Gauss J (1993) Effects of electron correlation in the calculation of nuclear-magnetic-resonance chemical-shifts. *J Chem Phys* 99(5):3629–3643
122. Gao HW et al (2010) Comparison of different theory models and basis sets in the calculations of structures and C-13 NMR spectra of [Pt(en)(CBDCA-O, O)], an analogue of the antitumor drug carboplatin. *J Phys Chem B* 114(11):4056–4062
123. Wu A et al (2007) Systematic studies on the computation of nuclear magnetic resonance shielding constants and chemical shifts: the density functional models. *J Comput Chem* 28(15):2431–2442
124. Giesen DJ, Zumbulyadis N (2002) A hybrid quantum mechanical and empirical model for the prediction of isotropic C-13 shielding constants of organic molecules. *Phys Chem Chem Phys* 4(22):5498–5507
125. Hoffmann F et al (2017) Improved quantum chemical NMR chemical shift prediction of metabolites in aqueous solution toward the validation of unknowns. *J Phys Chem A* 121(16):3071–3078
126. Aliev AE, Courtier-Murias D, Zhou S (2009) Scaling factors for carbon NMR chemical shifts obtained from DFF B3LYP calculations. *J Mol Struct Theochem* 893(1–3):1–5
127. Willoughby PH, Jansma MJ, Hoyer TR (2014) A guide to small-molecule structure assignment through computation of (H-1 and C-13) NMR chemical shifts. *Nat Protoc* 9(3):643–660
128. Pierens GK (2014) H-1 and C-13 NMR scaling factors for the calculation of chemical shifts in commonly used solvents using density functional theory. *J Comput Chem* 35(18):1388–1394
129. Caputo MC, Provasi PF, Sauer SPA (2018) The role of explicit solvent molecules in the calculation of NMR chemical shifts of glycine in water. *Theor Chem Accounts* 137(7):1–8
130. Feunang YD et al (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminformatics* 8:1–20
131. Yesiltepe Y et al (2018) An automated framework for NMR chemical shift calculations of small organic molecules. *J Cheminformatics* 10:1–16
132. Koster J, Rahmann S (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28(19):2520–2522
133. Weininger D (1988) Smiles, a chemical language and information-system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28(1):31–36
134. Valiev M et al (2010) NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. *Comput Phys Commun* 181(9):1477–1489
135. Lee CT, Yang WT, Parr RG (1988) Development of the colle-salvetti correlation-energy formula into a functional of the electron-density. *Phys Rev B* 37(2):785–789
136. Becke AD (1993) A new mixing of hartree-fock and local density-functional theories. *J Chem Phys* 98(2):1372–1377
137. Binkley JS, Pople JA, Hehre WJ (1980) Self-consistent molecular-orbital methods. 21. Small split-valence basis-sets for 1st-row elements. *J Am Chem Soc* 102(3):939–947
138. Gordon MS et al (1982) Self-consistent molecular-orbital methods. 22. Small split-valence basis-sets for 2nd-row elements. *J Am Chem Soc* 104(10):2797–2803
139. Schuchardt KL et al (2007) Basis set exchange: a community database for computational sciences. *J Chem Inf Model* 47(3):1045–1052

140. Saielli G et al (2011) Addressing the stereochemistry of complex organic molecules by density functional theory-NMR: vannusal B in retrospective. *J Am Chem Soc* 133(15):6072–6077
141. Tantillo DJ (2013) Walking in the woods with quantum chemistry—applications of quantum chemical calculations in natural products research. *Nat Prod Rep* 30(8):1079–1086
142. Klamt A, Schüürmann G (1993) COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J Chem Soc Perkin Trans 2*(5):799–805
143. Feller D (1996) The role of databases in support of computational chemistry calculations. *J Comput Chem* 17(13):1571–1586
144. Xin D et al (2017) Development of a ¹³C NMR chemical shift prediction procedure using B3LYP/cc-pVDZ and empirically derived systematic error correction terms: a computational small molecule structure elucidation method. *J Org Chem* 82(10):5135–5145
145. Ditchfield R (1974) Self-consistent perturbation-theory of diamagnetism. 1. Gauge-invariant Lcao method for Nmr chemical-shifts. *Mol Phys* 27(4):789–807
146. Oliveira FM et al (2016) Evaluation of some density functional methods for the estimation of hydrogen and carbon chemical shifts of phosphoramidates. *Comput Theor Chem* 1090:218–224
147. Smith SG, Goodman JM (2010) Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: the DP4 probability. *J Am Chem Soc* 132(37):12946–12959
148. Grimblat N, Zanardi MM, Sarotti AM (2015) Beyond DP4: an Improved probability for the stereochemical assignment of isomeric compounds using quantum chemical calculations of NMR shifts. *J Org Chem* 80(24):12526–12534
149. Navarro-Vazquez A (2017) State of the art and perspectives in the application of quantum chemical prediction of H-1 and C-13 chemical shifts and scalar couplings for structural elucidation of organic compounds. *Magn Reson Chem* 55(1):29–32
150. Ermanis K et al (2017) Doubling the power of DP4 for computational structure elucidation. *Org Biomol Chem* 15(42):8998–9007
151. Renslow RS et al (2014) A biofilm microreactor system for simultaneous electrochemical and nuclear magnetic resonance techniques. *Water Sci Technol* 69(5):966–973
152. Sutovich KJ et al (1999) Simultaneous quantification of Bronsted- and Lewis-acid sites in a USY zeolite. *J Catal* 183(1):155–158
153. Mueller LJ (1997) Chemical exchange in nuclear magnetic resonance. California Institute of Technology
154. Munkres J (1957) Algorithms for the assignment and transportation problems. *J Soc Ind Appl Math* 5(1):32–38
155. Kuhn HW (1955) The Hungarian method for the assignment problem. *Naval Res Logist Q* 2(1):83–97
156. Kuhn HW (1956) Variants of the Hungarian method for assignment problems. *Naval Res Logist Q* 3(4):253–258
157. Cui H, et al (2016) Solving large-scale assignment problems by Kuhn-Munkres algorithm. In: Proceedings of the 2nd international conference on advances in mechanical engineering and industrial informatics (Ameii 2016), vol 73, pp 822–827.
158. NaganaGowda GA, Raftery D (2017) Recent advances in NMR-based metabolomics. *Anal Chem* 89(1):490–510
159. Bingol K (2018) Recent advances in targeted and untargeted metabolomics by NMR and MS/NMR methods. *High Throughput* 7(2):9
160. Hogben HJ et al (2011) Spinach—a software library for simulation of spin dynamics in large spin systems. *J Magn Reson* 208(2):179–194
161. Bingol K et al (2015) Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. *Anal Chem* 87(7):3864–3870

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

