An Initial Investigation of Accuracy Required for the Identification of Small Molecules in Complex Samples Using Quantum Chemical Calculated NMR Chemical Shifts

Yasemin Yesiltepe

Pacific Northwest National Laboratory <https://orcid.org/0000-0001-9437-2347>

Niri Govind

Pacific Northwest National Laboratory <https://orcid.org/0000-0003-3625-366X>

Thomas O. Metz

Pacific Northwest National Laboratory <https://orcid.org/0000-0003-2448-6845>

Ryan Renslow (✉ ryan.renslow@pnnl.gov)

Pacific Northwest National Laboratory <https://orcid.org/0000-0002-3969-5570>

Research article

Keywords: metabolomics, unambiguous, nuclear magnetic resonance (NMR)

Posted Date: December 7th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-121004/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Title

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

Authors:

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

1 The Gene and Linda Voiland School of Chemical Engineering and Bioengineering, Washington State University, Pullman, WA, USA.

2 Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA

*Corresponding author: Ryan Renslow (ryan.renslow@pnnl.gov)

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 ^{13}C and ^1H NMR spectra with respect to experimentally measured values. This leads us to the question--what accuracy is required in predicted ^{13}C 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 ^{13}C chemical shifts in water are below 0.6 ppm and 7.1 ppm, respectively, and below 0.5 ppm 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%.

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 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 (FooDB) [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 [122, 123], (ii) the quality of the method is improved for geometry optimization [124, 125], (iii) a scaling procedure is employed [126, 127], (iv) conformational sampling is applied [128], and (v) solvation is taken into account appropriately [129, 130]. 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 [131]. 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 Supporting Information (SI).

Figure 1. Histograms depicting the number of molecules in each set for a given number of carbon atoms in a) Set I and b) Set II, and hydrogen atoms in c) Set I and d) Set II.

Computational details

The NMR chemical shifts for all molecules in this study were calculated using the *In Silico* Chemical Library Engine (ISiCLE) [132] (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 [133]. This pipeline takes SMILES [134] (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 [135] (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 [136, 137] with 3-21G basis set [138-140]. We chose this level of theory due to our available computational resources, particularly considering the treatments for the geometry optimization of over 11k 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 [128, 141, 142], 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 SI. The inclusion of solvent is via the COnductor-like Screening MOdel (COSMO) [143] solvation modeling. NMR isotropic shieldings were calculated for all optimized molecules at the B3LYP/cc-pVDZ [140, 144] level of theory. Based on our previous assessment [132], this method provides reliable chemical shifts [112] and yields isotropic shieldings with a reasonably low computational cost [145]. The gauge-invariant atomic orbital (GIAO) approach [146] was used to compute ^{13}C and ^1H NMR chemical shifts. The computed chemical shifts are provided in SI.

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) (Eqn 1), root mean square error (RMSE) (Eqn 2), corrected mean absolute error (CMAE) (Eqn 3), and correlation coefficients (e.g., the Pearson correlation coefficient).

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

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

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

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 Figure 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 [132, 147-150], we assume errors for both carbons and protons follow a Gaussian distribution [145, 151] 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 [145, 148]. In this study, σ is taken in the range of 0.5 to 50 ppm and 0.1 to 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 [152-154].

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 [155-157], which gives the minimum distance score (i.e. error) of two sets, within a feasible computational time bounded by a polynomial expression [158]. The Munkres algorithm minimizes the total error or summation of squared differences between each assignment. It is based on the following principle:

Let $S1$ and $S2$ 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 $S1$, b_j is the n^{th} element of $S2$ 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 [148], 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 SI.

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 SI.

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).

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

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.

Figure 3 Averaged percentages of molecules being identified within the first 1, 2, 5, and 10 candidate molecules at different Gaussian standard deviation (σ) values (ppm) for Set I and Set II when ^{13}C NMR chemical shifts are used alone.

Figure 4 Averaged percent of molecules being identified within the first 1, 2, 5, and 10 candidate molecules at different Gaussian standard deviation (σ) values (ppm) for Set I and Set II when ^1H NMR chemical shifts are used alone.

Figure 3 and **Figure 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 ppm 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 ppm 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 **Figure 3** and **Figure 4**, and the full list including 50-100% of identification is given in the SI.

Figure 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.

Error! Reference source not found. 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 ^{1}H 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 (Eqn 4) (see SI 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.

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

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 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 SI.

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

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 SI.

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 **Figure 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 **Figure 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 **Figure 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. [159, 160]. This is still, however, less than Case I by 84% and 93% for pairs and triplets, respectively (red line in Figure 6). Standard deviations of ranks and computational times of runs are given in the SI.

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.

Figure 6. Average ranks of Case I and Case II with known/unknown number of molecules in samples at optimum points for Set I (water soluble molecules).

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 [[161](#)]) 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 [162]. 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.

Declarations**Availability of data and materials**

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

Competing interests

The authors declare no competing interests.

Authors' contributions

Yasemin Yesiltepe performed all calculations, analyzed the results, created the figures, and was the primary author. Niri Govind assisted with the DFT calculations. Thomas O. Metz and Ryan S. Renslow oversaw the research, helped with analysis, and provided the funding for the study. All authors contributed to writing.

Funding and Acknowledgements

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.

References

1. German, J.B., B.D. Hammock, and S.M. Watkins, *Metabolomics: building on a century of biochemistry to guide human health*. Metabolomics, 2005. **1**(1): p. 3-9.
2. Wishart, D.S., *Current progress in computational metabolomics*. Brief Bioinform, 2007. **8**(5): p. 279-93.
3. Shulaev, V., *Metabolomics technology and bioinformatics*. Brief Bioinform, 2006. **7**(2): p. 128-39.
4. Kosmidis, A.K., et al., *Metabolomic fingerprinting: challenges and opportunities*. Crit Rev Biomed Eng, 2013. **41**(3): p. 205-21.
5. Nicholson, J.K. and I.D. Wilson, *Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism*. Nat Rev Drug Discov, 2003. **2**(8): p. 668-76.
6. Winnike, J.H., et al., *Use of pharmaco-metabonomics for early prediction of acetaminophen-induced hepatotoxicity in humans*. Clin Pharmacol Ther, 2010. **88**(1): p. 45-51.
7. Holmes, E., et al., *Human metabolic phenotype diversity and its association with diet and blood pressure*. Nature, 2008. **453**(7193): p. 396-400.
8. Beckonert, O., et al., *Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts*. Nature Protocols, 2007. **2**(11): p. 2692-2703.
9. Nicholson, J.K., J.C. Lindon, and E. Holmes, *'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data*. Xenobiotica, 1999. **29**(11): p. 1181-1189.
10. Nicholson, J.K., et al., *750 MHz ¹H and ¹H-¹³C NMR spectroscopy of human blood plasma*. Anal Chem, 1995. **67**(5): p. 793-811.
11. Smith, C.A., et al., *XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification*. Analytical Chemistry, 2006. **78**(3): p. 779-787.
12. Dettmer, K., P.A. Aronov, and B.D. Hammock, *Mass spectrometry-based metabolomics*. Mass Spectrometry Reviews, 2007. **26**(1): p. 51-78.
13. Want, E.J., B.F. Cravatt, and G. Siuzdak, *The expanding role of mass spectrometry in metabolite profiling and characterization*. Chembiochem, 2005. **6**(11): p. 1941-51.
14. Dunn, W.B., N.J. Bailey, and H.E. Johnson, *Measuring the metabolome: current analytical technologies*. Analyst, 2005. **130**(5): p. 606-25.
15. Hollywood, K., D.R. Brison, and R. Goodacre, *Metabolomics: Current technologies and future trends*. Proteomics, 2006. **6**(17): p. 4716-4723.
16. Moco, S., et al., *Metabolomics technologies and metabolite identification*. Trac-Trends in Analytical Chemistry, 2007. **26**(9): p. 855-866.
17. Smith, C.A., et al., *METLIN: a metabolite mass spectral database*. Ther Drug Monit, 2005. **27**(6): p. 747-51.
18. Wishart, D.S., et al., *HMDB 3.0--The Human Metabolome Database in 2013*. Nucleic Acids Res, 2013. **41**(Database issue): p. D801-7.
19. Ulrich, E.L., et al., *BioMagResBank*. Nucleic Acids Res, 2008. **36**(Database issue): p. D402-8.
20. Pence, H.E. and A. Williams, *ChemSpider: An Online Chemical Information Resource*. Journal of Chemical Education, 2010. **87**(11): p. 1123-1124.
21. Tautenhahn, R., et al., *XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data*. Analytical Chemistry, 2012. **84**(11): p. 5035-5039.
22. Williams, A.J., *A perspective of publicly accessible/open-access chemistry databases*. Drug Discov Today, 2008. **13**(11-12): p. 495-501.

23. Sitzmann, M., I.V. Filippov, and M.C. Nicklaus, *Internet resources integrating many small-molecule databases*. SAR QSAR Environ Res, 2008. **19**(1-2): p. 1-9.
24. Kutzler, F.W., et al., *Charge-Density and Bonding in (5,10,15,20-Tetramethylporphyrinato)Nickel(II) - a Combined Experimental and Theoretical-Study*. Journal of the American Chemical Society, 1983. **105**(10): p. 2996-3004.
25. Stimpson, D.I. and J.R. Cann, *A Combined Theoretical and Experimental-Study of the Interaction of Metrizamide with Proteins*. Archives of Biochemistry and Biophysics, 1981. **211**(1): p. 403-412.
26. Cripps, S.C., R.S. Orton, and J.E. Carroll, *Combined Theoretical and Experimental Studies of a Push-Pull Trapatt Circuit*. International Journal of Electronics, 1974. **37**(1): p. 1-21.
27. Gaulton, A., et al., *ChEMBL: a large-scale bioactivity database for drug discovery*. Nucleic Acids Res, 2012. **40**(Database issue): p. D1100-7.
28. Izgi, T., et al., *FT-IR and NMR investigation of 2-(1-cyclohexenyl)ethylamine: A combined experimental and theoretical study*. Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy, 2007. **68**(1): p. 55-62.
29. de Matos, P., et al., *Chemical Entities of Biological Interest: an update*. Nucleic Acids Research, 2010. **38**: p. D249-D254.
30. Kwan, E.E. and R.Y. Liu, *Enhancing NMR Prediction for Organic Compounds Using Molecular Dynamics*. Journal of Chemical Theory and Computation, 2015. **11**(11): p. 5083-5089.
31. Knox, C., et al., *DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs*. Nucleic Acids Research, 2011. **39**: p. D1035-D1041.
32. Ulrich, E.L., et al., *BioMagResBank*. Nucleic Acids Research, 2008. **36**: p. D402-D408.
33. Wishart, D.S., et al., *HMDB: a knowledgebase for the human metabolome*. Nucleic Acids Res, 2009. **37**(Database issue): p. D603-10.
34. Blum, L.C. and J.L. Reymond, *970 million druglike small molecules for virtual screening in the chemical universe database GDB-13*. J Am Chem Soc, 2009. **131**(25): p. 8732-3.
35. Jewison, T., et al., *SMPDB 2.0: big improvements to the Small Molecule Pathway Database*. Nucleic Acids Res, 2014. **42**(Database issue): p. D478-84.
36. Frolkis, A., et al., *SMPDB: The Small Molecule Pathway Database*. Nucleic Acids Res, 2010. **38**(Database issue): p. D480-7.
37. Richard, A.M. and C.R. Williams, *Distributed structure-searchable toxicity (DSSTox) public database network: a proposal*. Mutat Res, 2002. **499**(1): p. 27-52.
38. Guo, A.C., et al., *ECMDB: the E. coli Metabolome Database*. Nucleic Acids Res, 2013. **41**(Database issue): p. D625-30.
39. Sajed, T., et al., *ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli*. Nucleic Acids Res, 2016. **44**(D1): p. D495-501.
40. Keseler, I.M., et al., *The EcoCyc database: reflecting new knowledge about Escherichia coli K-12*. Nucleic Acids Research, 2017. **45**(D1): p. D543-D550.
41. Scalbert, A., et al., *Databases on Food Phytochemicals and Their Health-Promoting Effects*. Journal of Agricultural and Food Chemistry, 2011. **59**(9): p. 4331-4348.
42. Fahy, E., et al., *Update of the LIPID MAPS comprehensive classification system for lipids*. Journal of Lipid Research, 2009. **50**: p. S9-S14.
43. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes*. Nucleic Acids Research, 2018. **46**(D1): p. D633-D639.
44. MolMall. [cited 2019 8/1]; Available from: <http://www.molmall.net/>.
45. Banerjee, P., et al., *Super Natural II-a database of natural products*. Nucleic Acids Research, 2015. **43**(D1): p. D935-D939.
46. Wishart, D., et al., *T3DB: the toxic exposome database*. Nucleic Acids Res, 2015. **43**(Database issue): p. D928-34.

47. Lim, E., et al., *T3DB: a comprehensively annotated database of common toxins and their targets*. Nucleic Acids Research, 2010. **38**: p. D781-D786.
48. Richard, A.M., et al., *ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology*. Chemical Research in Toxicology, 2016. **29**(8): p. 1225-1251.
49. Gu, J.Y., et al., *Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology*. Plos One, 2013. **8**(4).
50. Sterling, T. and J.J. Irwin, *ZINC 15-Ligand Discovery for Everyone*. Journal of Chemical Information and Modeling, 2015. **55**(11): p. 2324-2337.
51. Wishart, D.S., *Advances in metabolite identification*. Bioanalysis, 2011. **3**(15): p. 1769-1782.
52. Xiao, J.F., B. Zhou, and H.W. Ransom, *Metabolite identification and quantitation in LC-MS/MS-based metabolomics*. Trac-Trends in Analytical Chemistry, 2012. **32**: p. 1-14.
53. NIST 17 MS/MS Library. [cited 2019 05.01]; Available from: <https://www.sisweb.com/software/nist-msms.htm>.
54. The NIST 17 Mass Spectral Library. June 2017 [cited 2019 05.01]; Available from: <https://www.sisweb.com/software/ms/nist.htm#stats>.
55. The Human Metabolome Library (HML). [cited 2019 05.01]; Available from: <http://www.hmdb.ca/hml>.
56. Wishart, D.S., et al., *HMDB 4.0: the human metabolome database for 2018*. Nucleic Acids Res, 2018. **46**(D1): p. D608-D617.
57. Wishart, D.S., et al., *HMDB: the Human Metabolome Database*. Nucleic Acids Res, 2007. **35**(Database issue): p. D521-6.
58. ZINC 15, a free database of commercially-available compounds. [cited 2019 05.01]; Available from: <http://zinc15.docking.org/>.
59. Sterling, T. and J.J. Irwin, *ZINC 15--Ligand Discovery for Everyone*. J Chem Inf Model, 2015. **55**(11): p. 2324-37.
60. Styczynski, M.P., et al., *Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery*. Anal Chem, 2007. **79**(3): p. 966-73.
61. Staniek, A., H.J. Woerdenbag, and O. Kayser, *Endophytes: exploiting biodiversity for the improvement of natural product-based drug discovery*. Journal of Plant Interactions, 2008. **3**(2): p. 75-93.
62. Tulp, M. and L. Bohlin, *Functional versus chemical diversity: is biodiversity important for drug discovery?* Trends in Pharmacological Sciences, 2002. **23**(5): p. 225-231.
63. Sumner, L.W., et al., *Proposed minimum reporting standards for chemical analysis*. Metabolomics, 2007. **3**(3): p. 211-221.
64. DeHaven, C.D., et al., *Organization of GC/MS and LC/MS metabolomics data into chemical libraries*. Journal of Cheminformatics, 2010. **2**.
65. Dobson, C.M., *Chemical space and biology*. Nature, 2004. **432**(7019): p. 824-828.
66. Patti, G.J., et al., *A View from Above: Cloud Plots to Visualize Global Metabolomic Data*. Analytical Chemistry, 2013. **85**(2): p. 798-804.
67. Weckwerth, W., *Metabolomics in systems biology*. Annual Review of Plant Biology, 2003. **54**: p. 669-689.
68. Salek, R.M., et al., *The role of reporting standards for metabolite annotation and identification in metabolomic studies*. Gigascience, 2013. **2**.
69. Fiehn, O., et al., *The metabolomics standards initiative (MSI)*. Metabolomics, 2007. **3**(3): p. 175-178.
70. Beisken, S., M. Eiden, and R.M. Salek, *Getting the right answers: understanding metabolomics challenges*. Expert Review of Molecular Diagnostics, 2015. **15**(1): p. 97-109.

71. Di Stefano, V., et al., *Applications of liquid chromatography-mass spectrometry for food analysis*. Journal of Chromatography A, 2012. **1259**: p. 74-85.
72. Garcia, A. and C. Barbas, *Gas chromatography-mass spectrometry (GC-MS)-based metabolomics*. Methods Mol Biol, 2011. **708**: p. 191-204.
73. Schymanski, E.L., et al., *Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence*. Environmental Science & Technology, 2014. **48**(4): p. 2097-2098.
74. Tang, H.R., et al., *Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma*. Analytical Biochemistry, 2004. **325**(2): p. 260-272.
75. Nicholson, J.K. and I.D. Wilson, *Understanding 'global' systems biology: Metabonomics and the continuum of metabolism*. Nature Reviews Drug Discovery, 2003. **2**(8): p. 668-676.
76. Kangas, L.J., et al., *In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids*. Bioinformatics, 2012. **28**(13): p. 1705-1713.
77. Allen, F., R. Greiner, and D. Wishart, *Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification*. Metabolomics, 2015. **11**(1): p. 98-110.
78. Wolf, S., et al., *In silico fragmentation for computer assisted identification of metabolite mass spectra*. BMC Bioinformatics, 2010. **11**.
79. Bouteiller, Y., et al., *Transferable Specific Scaling Factors for Interpretation of Infrared Spectra of Biomolecules from Density Functional Theory*. Journal of Physical Chemistry A, 2008. **112**(46): p. 11656-11660.
80. Colby, S.M., et al., *ISICLE: A quantum chemistry pipeline for establishing in silico collision cross section libraries*. Analytical chemistry, 2019. **91**(7): p. 4346-4356.
81. Nuñez, J.R., et al., *Advancing Standards-Free Methods for the Identification of Small Molecules in Complex Samples*. arXiv preprint arXiv:1810.07367, 2018.
82. Casabianca, L.B. and A.C. De Dios, *Ab initio calculations of NMR chemical shifts*. Journal of Chemical Physics, 2008. **128**(5).
83. Lodewyk, M.W., M.R. Siebert, and D.J. Tantillo, *Computational prediction of ¹H and ¹³C chemical shifts: a useful tool for natural product, mechanistic, and synthetic organic chemistry*. Chem Rev, 2012. **112**(3): p. 1839-62.
84. Hill, D.E., N. Vasdev, and J.P. Holland, *Evaluating the accuracy of density functional theory for calculating ¹H-1 and ¹³C-13 NMR chemical shifts in drug molecules*. Computational and Theoretical Chemistry, 2015. **1051**: p. 161-172.
85. Lomas, J.S., *¹H-1 NMR spectra of alcohols in hydrogen bonding solvents: DFT/GIAO calculations of chemical shifts*. Magnetic Resonance in Chemistry, 2016. **54**(1): p. 28-38.
86. Zheng, X.Y., et al., *Structural Elucidation of cis/trans Dicafeoylquinic Acid Photoisomerization Using Ion Mobility Spectrometry-Mass Spectrometry*. Journal of Physical Chemistry Letters, 2017. **8**(7): p. 1381-1388.
87. Zheng, X.Y., et al., *Enhancing glycan isomer separations with metal ions and positive and negative polarity ion mobility spectrometry-mass spectrometry analyses*. Analytical and Bioanalytical Chemistry, 2017. **409**(2): p. 467-476.
88. Nunez, J.R., et al., *Evaluation of In Silico Multi-Feature Libraries for Providing Evidence for the Presence of Small Molecules in Synthetic Blinded Samples*. Journal of Chemical Information and Modeling, 2019.
89. Forsyth, D.A. and A.B. Seabag, *Computed ¹³C NMR chemical shifts via empirically scaled GIAO shieldings and molecular mechanics geometries. Conformation and configuration from ¹³C shifts*. Journal of the American Chemical Society, 1997. **119**(40): p. 9483-9494.
90. Auer, A.A., J. Gauss, and J.F. Stanton, *Quantitative prediction of gas-phase ¹³C nuclear magnetic shielding constants*. Journal of Chemical Physics, 2003. **118**(23): p. 10407-10417.

91. Mothana, B., F.Q. Ban, and R.J. Boyd, *Validation of a computational scheme to study N-15 and C-13 nuclear shielding constants*. Chemical Physics Letters, 2005. **401**(1-3): p. 7-12.
92. Saito, H., *Conformation-Dependent C-13 Chemical-Shifts - a New Means of Conformational Characterization as Obtained by High-Resolution Solid-State C-13 Nmr*. Magnetic Resonance in Chemistry, 1986. **24**(10): p. 835-852.
93. Jaime, C., et al., *C-13 Nmr Chemical-Shifts - a Single Rule to Determine the Conformation of Calix[4]Arenes*. Journal of Organic Chemistry, 1991. **56**(10): p. 3372-3376.
94. Yannoni, C.S., et al., *C-13 Nmr-Study of the C60 Cluster in the Solid-State - Molecular-Motion and Carbon Chemical-Shift Anisotropy*. Journal of Physical Chemistry, 1991. **95**(1): p. 9-10.
95. Malkin, V.G., et al., *Solvent effect on the NMR chemical shieldings in water calculated by a combination of molecular dynamics and density functional theory*. Chemistry-a European Journal, 1996. **2**(4): p. 452-457.
96. Casanovas, J., et al., *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*. Journal of Organic Chemistry, 2001. **66**(11): p. 3775-3782.
97. Benzi, C., et al., *Reliable NMR chemical shifts for molecules in solution by methods rooted in density functional theory*. Magnetic Resonance in Chemistry, 2004. **42**: p. S57-S67.
98. Kiamco, M.M., et al., *Structural and metabolic responses of Staphylococcus aureus biofilms to hyperosmotic and antibiotic stress*. Biotechnol Bioeng, 2018. **115**(6): p. 1594-1603.
99. Dreyer, D.R., et al., *Elucidating the structure of poly(dopamine)*. Langmuir, 2012. **28**(15): p. 6428-35.
100. Xin, D.Y., et al., *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*. Journal of Organic Chemistry, 2017. **82**(10): p. 5135-5145.
101. Garcellano, R.C., et al., *Isolation of Tryptanthrin and Reassessment of Evidence for Its Isobaric Isostere Wrightiadione in Plants of the Wrightia Genus*. J Nat Prod, 2018.
102. Kutateladze, A.G. and D.S. Reddy, *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, 2017. **82**(7): p. 3368-3381.
103. Kutateladze, A.G., E.H. Krenske, and C.M. Williams, *Reassignments and Corroborations of Oxo-Bridged Natural Products Directed by OSE and DU8+ NMR Computation*. Angew Chem Int Ed Engl, 2019. **58**(21): p. 7107-7112.
104. Jain, R., T. Bally, and P.R. Rablen, *Calculating accurate proton chemical shifts of organic molecules with density functional methods and modest basis sets*. J Org Chem, 2009. **74**(11): p. 4017-23.
105. Perez, M., et al., *Accuracy vs time dilemma on the prediction of NMR chemical shifts: A case study (chloropyrimidines)*. Journal of Organic Chemistry, 2006. **71**(8): p. 3103-3110.
106. Barone, G., et al., *Determination of the relative stereochemistry of flexible organic compounds by ab initio methods: Conformational analysis and Boltzmann-averaged GIAO C-13 NMR chemical shifts*. Chemistry-a European Journal, 2002. **8**(14): p. 3240-3245.
107. Barone, G., et al., *Structure validation of natural products by quantum-mechanical GIAO calculations of C-13 NMR chemical shifts*. Chemistry-a European Journal, 2002. **8**(14): p. 3233-3239.
108. Remya, K. and C.H. Suresh, *Which density functional is close to CCSD accuracy to describe geometry and interaction energy of small non-covalent dimers? A benchmark study using gaussian09*. Journal of Computational Chemistry, 2013. **34**(15): p. 1341-1353.
109. Zhao, Y. and D.G. Truhlar, *Improved description of nuclear magnetic resonance chemical shielding constants using the M06-L meta-generalized-gradient-approximation density functional*. Journal of Physical Chemistry A, 2008. **112**(30): p. 6794-6799.

110. Magyarfalvi, G. and P. Pulay, *Assessment of density functional methods for nuclear magnetic resonance shielding calculations*. Journal of Chemical Physics, 2003. **119**(3): p. 1350-1357.
111. Cimino, P., et al., *Comparison of different theory models and basis sets in the calculation of C-13 NMR chemical shifts of natural products*. Magnetic Resonance in Chemistry, 2004. **42**: p. S26-S33.
112. Tormena, C.F. and G.V.J. da Silva, *Chemical shifts calculations on aromatic systems: a comparison of models and basis sets*. Chemical Physics Letters, 2004. **398**(4-6): p. 466-470.
113. Cramer, C.J. and D.G. Truhlar, *Implicit solvation models: Equilibria, structure, spectra, and dynamics*. Chemical Reviews, 1999. **99**(8): p. 2161-2200.
114. Wiitala, K.W., T.R. Hoyer, and C.J. Cramer, *Hybrid density functional methods empirically optimized for the computation of C-13 and H-1 chemical shifts in chloroform solution*. Journal of Chemical Theory and Computation, 2006. **2**(4): p. 1085-1092.
115. Reddy, G. and A. Yethiraj, *Implicit and explicit solvent models for the simulation of dilute polymer solutions*. Macromolecules, 2006. **39**(24): p. 8536-8542.
116. Smirnov, S.N., et al., *Hydrogen deuterium isotope effects on the NMR chemical shifts and geometries of intermolecular low-barrier hydrogen-bonded complexes*. Journal of the American Chemical Society, 1996. **118**(17): p. 4094-4101.
117. Benedict, H., et al., *Hydrogen/deuterium isotope effects on the N-15 NMR chemical shifts and geometries of low-barrier hydrogen bonds in the solid state*. Journal of Molecular Structure, 1996. **378**(1): p. 11-16.
118. Gidley, M.J. and S.M. Bociek, *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*. Journal of the American Chemical Society, 1988. **110**(12): p. 3820-3829.
119. Buckingham, A.D., *Chemical Shifts in the Nuclear Magnetic Resonance Spectra of Molecules Containing Polar Groups*. Canadian Journal of Chemistry-Revue Canadienne De Chimie, 1960. **38**(2): p. 300-307.
120. Osmialowski, B., E. Kolehmainen, and R. Gawinecki, *GIAO/DFT calculated chemical shifts of tautomeric species. 2-Phenacylpyridines and (Z)-2-(2-hydroxy-2-phenylvinyl)pyridines*. Magnetic Resonance in Chemistry, 2001. **39**(6): p. 334-340.
121. Gauss, J., *Effects of Electron Correlation in the Calculation of Nuclear-Magnetic-Resonance Chemical-Shifts*. Journal of Chemical Physics, 1993. **99**(5): p. 3629-3643.
122. Jain, R., T. Bally, and P.R. Rablen, *Calculating Accurate Proton Chemical Shifts of Organic Molecules with Density Functional Methods and Modest Basis Sets*. Journal of Organic Chemistry, 2009. **74**(11): p. 4017-4023.
123. Gao, H.W., et al., *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*. Journal of Physical Chemistry B, 2010. **114**(11): p. 4056-4062.
124. Wu, A., et al., *Systematic studies on the computation of nuclear magnetic resonance shielding constants and chemical shifts: The density functional models*. Journal of Computational Chemistry, 2007. **28**(15): p. 2431-2442.
125. Giesen, D.J. and N. Zumbulyadis, *A hybrid quantum mechanical and empirical model for the prediction of isotropic C-13 shielding constants of organic molecules*. Physical Chemistry Chemical Physics, 2002. **4**(22): p. 5498-5507.
126. Hoffmann, F., et al., *Improved Quantum Chemical NMR Chemical Shift Prediction of Metabolites in Aqueous Solution toward the Validation of Unknowns*. J Phys Chem A, 2017. **121**(16): p. 3071-3078.

127. Aliev, A.E., D. Courtier-Murias, and S. Zhou, *Scaling factors for carbon NMR chemical shifts obtained from DFF B3LYP calculations*. Journal of Molecular Structure-Theochem, 2009. **893**(1-3): p. 1-5.
128. Willoughby, P.H., M.J. Jansma, and T.R. Hoyer, *A guide to small-molecule structure assignment through computation of (¹H- and ¹³C-) NMR chemical shifts*. Nature Protocols, 2014. **9**(3): p. 643-660.
129. Pierens, G.K., *H-1 and C-13 NMR Scaling Factors for the Calculation of Chemical Shifts in Commonly Used Solvents Using Density Functional Theory*. Journal of Computational Chemistry, 2014. **35**(18): p. 1388-1394.
130. Caputo, M.C., P.F. Provasi, and S.P.A. Sauer, *The role of explicit solvent molecules in the calculation of NMR chemical shifts of glycine in water*. Theoretical Chemistry Accounts, 2018. **137**(7).
131. Feunang, Y.D., et al., *ClassyFire: automated chemical classification with a comprehensive, computable taxonomy*. Journal of Cheminformatics, 2016. **8**.
132. Yesiltepe, Y., et al., *An automated framework for NMR chemical shift calculations of small organic molecules*. Journal of Cheminformatics, 2018. **10**.
133. Koster, J. and S. Rahmann, *Snakemake-a scalable bioinformatics workflow engine*. Bioinformatics, 2012. **28**(19): p. 2520-2522.
134. Weininger, D., *Smiles, a Chemical Language and Information-System .1. Introduction to Methodology and Encoding Rules*. Journal of Chemical Information and Computer Sciences, 1988. **28**(1): p. 31-36.
135. Valiev, M., et al., *NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations*. Computer Physics Communications, 2010. **181**(9): p. 1477-1489.
136. Lee, C.T., W.T. Yang, and R.G. Parr, *Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density*. Physical Review B, 1988. **37**(2): p. 785-789.
137. Becke, A.D., *A New Mixing of Hartree-Fock and Local Density-Functional Theories*. Journal of Chemical Physics, 1993. **98**(2): p. 1372-1377.
138. Binkley, J.S., J.A. Pople, and W.J. Hehre, *Self-Consistent Molecular-Orbital Methods .21. Small Split-Valence Basis-Sets for 1st-Row Elements*. Journal of the American Chemical Society, 1980. **102**(3): p. 939-947.
139. Gordon, M.S., et al., *Self-Consistent Molecular-Orbital Methods .22. Small Split-Valence Basis-Sets for 2nd-Row Elements*. Journal of the American Chemical Society, 1982. **104**(10): p. 2797-2803.
140. Schuchardt, K.L., et al., *Basis set exchange: A community database for computational sciences*. Journal of Chemical Information and Modeling, 2007. **47**(3): p. 1045-1052.
141. Saielli, G., et al., *Addressing the stereochemistry of complex organic molecules by density functional theory-NMR: vannusall B in retrospective*. J Am Chem Soc, 2011. **133**(15): p. 6072-7.
142. Tantillo, D.J., *Walking in the woods with quantum chemistry - applications of quantum chemical calculations in natural products research*. Natural Product Reports, 2013. **30**(8): p. 1079-1086.
143. Klamt, A. and G. Schuurmann, *COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient*. Journal of the Chemical Society, Perkin Transactions 2, 1993(5): p. 799-805.
144. Feller, D., *The role of databases in support of computational chemistry calculations*. Journal of Computational Chemistry, 1996. **17**(13): p. 1571-1586.
145. Xin, D., et al., *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, 2017. **82**(10): p. 5135-5145.
146. Ditchfield, R., *Self-Consistent Perturbation-Theory of Diamagnetism .1. Gauge-Invariant Lcao Method for Nmr Chemical-Shifts*. Molecular Physics, 1974. **27**(4): p. 789-807.

147. Oliveira, F.M., et al., *Evaluation of some density functional methods for the estimation of hydrogen and carbon chemical shifts of phosphoramidates*. Computational and Theoretical Chemistry, 2016. **1090**: p. 218-224.
148. Smith, S.G. and J.M. Goodman, *Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: the DP4 probability*. J Am Chem Soc, 2010. **132**(37): p. 12946-59.
149. Grimblat, N., M.M. Zanardi, and A.M. Sarotti, *Beyond DP4: an Improved Probability for the Stereochemical Assignment of Isomeric Compounds using Quantum Chemical Calculations of NMR Shifts*. J Org Chem, 2015. **80**(24): p. 12526-34.
150. Navarro-Vazquez, A., *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*. Magnetic Resonance in Chemistry, 2017. **55**(1): p. 29-32.
151. Ermanis, K., et al., *Doubling the power of DP4 for computational structure elucidation*. Org Biomol Chem, 2017. **15**(42): p. 8998-9007.
152. Renslow, R.S., et al., *A biofilm microreactor system for simultaneous electrochemical and nuclear magnetic resonance techniques*. Water Sci Technol, 2014. **69**(5): p. 966-73.
153. Sutovich, K.J., et al., *Simultaneous quantification of Bronsted- and Lewis-acid sites in a USY zeolite*. Journal of Catalysis, 1999. **183**(1): p. 155-158.
154. Mueller, L.J., *Chemical Exchange in Nuclear Magnetic Resonance*. 1997, California Institute of Technology.
155. Munkres, J., *Algorithms for the Assignment and Transportation Problems*. Journal of the Society for Industrial and Applied Mathematics, 1957. **5**(1): p. 32-38.
156. Kuhn, H.W., *The Hungarian Method for the assignment problem*. Naval Research Logistics Quarterly, 1955. **2**(1): p. 83-97.
157. Kuhn, H.W., *Variants of the hungarian method for assignment problems*. Naval Research Logistics Quarterly, 1956. **3**(4): p. 253-258.
158. Cui, H., et al., *Solving large-scale assignment problems by Kuhn-Munkres algorithm*. Proceedings of the 2nd International Conference on Advances in Mechanical Engineering and Industrial Informatics (Ameii 2016), 2016. **73**: p. 822-827.
159. Nagana Gowda, G.A. and D. Raftery, *Recent Advances in NMR-Based Metabolomics*. Anal Chem, 2017. **89**(1): p. 490-510.
160. Bingol, K., *Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods*. High Throughput, 2018. **7**(2).
161. Hogben, H.J., et al., *Spinach--a software library for simulation of spin dynamics in large spin systems*. J Magn Reson, 2011. **208**(2): p. 179-94.
162. Bingol, K., et al., *Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures*. Anal Chem, 2015. **87**(7): p. 3864-70.

Figures

Figure 1

Histograms depicting the number of molecules in each set for a given number of carbon atoms in a) Set I and b) Set II, and hydrogen atoms in c) Set I and d) Set II.

Figure 2

Flowchart for the identification of the compounds in a mixture.

Figure 3

Averaged percentages of molecules being identified within the first 1, 2, 5, and 10 candidate molecules at different Gaussian standard deviation (σ) values (ppm) for Set I and Set II when ^{13}C NMR chemical shifts are used alone.

Figure 4

Averaged percent of molecules being identified within the first 1,2,5, and 10 candidate molecules at different Gaussian standard deviation (σ) values (ppm) for Set I and Set II when ^1H NMR chemical shifts are used alone.

Figure 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.

Figure 6

Average ranks of Case I and Case II with known/unknown number of molecules in samples at optimum points for Set I (water soluble molecules).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DFTOutputPlaceholder.txt](#)
- [Pythonscripts.zipfile](#)
- [SetIIChloroformSoluableMolecules.xlsx](#)
- [SetIIChloroformSolubleCompoundsMOLfiles.zipfile](#)
- [SetIWaterSoluableMolecules.xlsx](#)
- [SetIWaterSolubleCompoundsMOLfiles.zipfile](#)
- [YesiltepeetalSI.docx](#)