



HHS Public Access

Author manuscript

J Phys Chem A. Author manuscript; available in PMC 2018 April 27.

Published in final edited form as:

J Phys Chem A. 2017 April 27; 121(16): 3071–3078. doi:10.1021/acs.jpca.7b01954.

Improved Quantum Chemical NMR Chemical Shift Prediction of Metabolites in Aqueous Solution Toward the Validation of Unknowns

Felix Hoffmann¹, Da-Wei Li², Daniel Sebastiani¹, and Rafael Brüschweiler^{2,3,4,*}

¹Institute of Chemistry, Martin-Luther-University Halle-Wittenberg, von-Danckelmann-Platz 4, 06120 Halle, Germany

²Campus Chemical Instrument Center, The Ohio State University, Columbus, Ohio 43210, United States

³Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States

⁴Department of Biological Chemistry and Pharmacology, The Ohio State University, Columbus, Ohio 43210, United States

Abstract

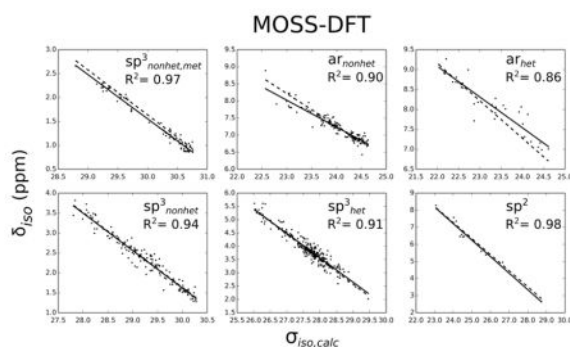
A quantum-chemistry based protocol, termed MOSS-DFT, is presented for the prediction of ¹³C and ¹H NMR chemical shifts of a wide range of organic molecules in aqueous solution, including metabolites. Molecular motif-specific linear scaling parameters are reported for five different density functional theory (DFT) methods (B97-2/pcS-1, B97-2/pcS-2, B97-2/pcS-3, B3LYP/pcS-2 and BLYP/pcS-2), which were applied to a large set of 176 metabolite molecules. The chemical shift root-mean-square deviations (RMSD) for the best method, B97-2/pcS-3, are 1.93 ppm and 0.154 ppm for ¹³C and ¹H chemical shifts, respectively. Excellent results have been obtained for chemical shifts of methyl and aromatic ¹³C and ¹H that are not directly bonded to a heteroatom (O, N, S, or P) with RMSD values of 1.15/0.079 ppm and 1.31/0.118 ppm, respectively. This study not only demonstrates how NMR chemical shift predictions in aqueous environment can be improved over the commonly used global linear scaling approach, but also allows for motif-specific error estimates, which are useful for an improved chemical shift-based verification of metabolite candidates of metabolomics samples containing unknown components.

Graphical Abstract

*To whom correspondence should be addressed: Rafael Brüschweiler, Ph.D., Department of Chemistry and Biochemistry, The Ohio State University, Columbus, Ohio 43210, bruschiweil.1@osu.edu.

5. Supporting Information

Histograms of absolute deviations, linear regression parameters for employed DFT methods, linear regression parameters and RMSD values of cross-validation analysis, list of molecule IDs used in the test set, table with molecules IDs and structures of the full MOSS-DFT set



1. Introduction

The analysis of complex metabolic mixtures of a wide range of biological systems using NMR spectroscopy has become increasingly popular over the past decade in the context of the rapidly growing field of metabolomics.^{1–4} This is in part due to the ability to simultaneously provide detailed spectroscopic information on many different metabolites in the same sample. In that way, new insights into the state of biological systems as well as into metabolic pathways are possible.^{5–8}

The concomitant development of databases and web-based query tools, like the Complex Mixture Analysis by NMR (COLMAR) database⁹, The Human Metabolome Database (HMDB)¹⁰ and the Biological Magnetic Resonance Data Bank (BMRB)¹¹, has further increased the usefulness of NMR spectroscopy and enabled the automated assignment of metabolites. However, the identification of unknown metabolites, which are metabolites that give rise to signals in the NMR spectra, but that have not been identified previously or are not part of commonly accessible NMR databases, remains a key challenge. In this regard, empirical chemical shift predictors, such as NMRPredict¹² as used by the MNova software,¹³ or the ACD/NMR predictor¹⁴ are useful to efficiently compare experimental chemical shift information with predicted chemical shifts of a large number of candidate structures, but their accuracy is determined by the nature and size of the underlying database.

As a consequence, one often encounters excellent prediction results for structures that are part of the underlying database, but structures (or substructures) that have not been considered during the fitting procedure are usually predicted with significantly reduced accuracy. Therefore, predictors that are less dependent on or even entirely independent of databases, potentially allow for the more balanced and more accurate chemical shift prediction for both known and unknown structures. In this context, quantum-chemical (QC) methods offer a promising alternative. However, their computational cost is in most cases several orders of magnitude larger and they require the use of high performance computing resources when a large number of predictions is needed. To keep the computational costs manageable, several approximations are usually employed, both regarding the system under study and the QC method itself. Although other procedures have been put forward in the literature,^{15–16} a widely-used approach is to calculate shielding constants only for a limited number of relevant conformers and to account for solvent effects implicitly. Among the

various QC methods, density functional theory (DFT) has been shown to yield both the accuracy and numerical efficiency to allow for NMR chemical shift calculations on a routine basis.^{17–24} Nevertheless, recent developments in wave function-based methods, especially MP2, offer potentially useful alternatives for small to medium-sized molecules.^{25–27}

The calculated shielding constants can be converted to chemical shifts in three ways: (i) by subtracting them from the shielding constant of an internal reference, such as tetramethylsilane (TMS) or 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS), (ii) by intermediate references or multi-standards, and (iii) by means of (linear) regression. Whereas the first two approaches usually rely on (often fortuitous) error cancellation between shielding constants calculated for the query and reference atom, linear regression allows for the correction of systematic errors as a function of the shielding constant. Several studies have reported excellent agreements with experimental shifts by using DFT together with linear regression and provided fitting parameters for several exchange-correlation/basis set combinations^{20, 28–30}, where some of them used multiple motif-dependent standards.³¹ However, most of them were conducted in organic solvents, such as chloroform, for a relatively small number of molecules. While these results are helpful in the field of general organic chemistry, they are somewhat less practical for metabolomics studies, because NMR chemical shifts are usually measured in aqueous solution. Naturally, the presence of water poses additional challenges for the prediction of NMR chemical shifts of metabolites, especially if combined with implicit solvent models. For example, chemical shifts of atoms in close proximity to hydrogen bond donors or acceptors will have different systematic errors than shifts of those atoms that are further away.³² Moreover, molecules that allow for intramolecular hydrogen bonding often exhibit large chemical shift deviations when implicit solvent models are employed. This is because conformational search and geometry optimization in implicit solvent favor geometries that are stabilized by intramolecular hydrogen bonds, which often leads to a biased conformational ensemble compared to the one in explicit water. As a result, atoms in different microscopic solvation environments will have different systematic errors.

In the present work, we report a DFT-based protocol, termed MOlecular motif-Specific Scaling of Density-Functional-Theory-based chemical shifts (MOSS-DFT), for NMR chemical shift predictions, which is based on a set of 176 molecules that are relevant for metabolomics studies. Chemical shielding constants are converted to chemical shifts by a motif-specific linear regression approach of calculated shielding constants to experimental chemical shift measured with 2D ¹H-¹³C HSQC spectroscopy. In the first part, details about the MOSS-DFT database construction and the computational methods are given. Thereafter, we provide the definition of motifs and compare our prediction results with those from earlier studies that employ global linear regressions. In the last part, we discuss the influence of basis set size and the exchange-correlation functional by comparing the results to various functional-basis set combinations.

2. Methods

For the MOSS-DFT database construction, a number of molecules were randomly picked from the COLMAR database. We used Open Babel³³ to generate 3D coordinates from 2D

structures and to adjust the protonation state of ionizable groups to pH 7, that is, all carboxylic, phosphonic and sulfonic acid groups were deprotonated and all amino groups were protonated. For the calculation of the chemical shifts we largely follow the approach described previously.²¹ All molecules were subjected to a conformational search, which was conducted with the MacroModel program³⁴ that is part of the Schrödinger suite. The OPLS 2005 force field³⁵ was used together with an implicit solvent model for water.³⁴ For sampling, the Monte Carlo Multiple Minimum (MCM) algorithm was chosen^{36–37} with the maximum step number set to 5000. Other program options have been set to their default values or set automatically by MacroModel, which is part of the automatic setup. To avoid unphysical conformational ensembles due to the implicit solvation model, the resulting conformers were checked for intramolecular hydrogen bonds, defined by a donor (D) – acceptor (A) distance of less than 3.5 Å and a D-H-A angle of $180^\circ \pm 30^\circ$. In the case of intramolecular hydrogen bonding in at least one of the conformers, the compound was removed from the database. In addition, molecules with long carbon hydride chains were excluded, if the conformational search yielded a large number (>100) of similarly low energy structures. The conformers obtained from the conformational search were further optimized at the DFT level using the Gaussian 09 program.³⁸ The B3LYP³⁹ exchange correlation functional was used together with the D3 dispersion correction⁴⁰ and the def2-TZVP basis set.⁴¹ In order to capture solvation effects, the conductor polarized continuum model (CPCM) was adopted using water as solvent.^{42–43} To ensure well-converged local minimum geometries, extremely tight convergence criteria have been used along with an ultrafine integration grid.

Convergence to local minimum structures has been additionally monitored by normal mode analysis at the same level of theory. In cases where two initially different conformers had essentially identical total energies ($E \approx 10^{-9}$ Hartree), it was assumed that the DFT optimization led to the same minimum structure and calculations were continued only for one of the structures. The conformer population p_i at 298.15 K was estimated from a Boltzmann analysis

$$p_i = \frac{e^{-E_i/RT}}{\sum_{j=1}^N e^{-E_j/RT}} \quad (1)$$

where E_i denotes the relative free energy (electronic + thermal free energy) of conformer i as estimated from the thermochemical analysis in Gaussian 09 with respect to the most stable conformer. RT is the product of the ideal gas constant and the absolute temperature, respectively, and N is the total number of distinct conformers as obtained from geometry optimization. NMR shielding constants were calculated based on the gauge-independent atomic orbitals (GIAO) approach as implemented in Gaussian 09.^{44–48} We calculated NMR shielding constants at the DFT level employing 5 different functional-basis set combinations (see Table 3). The computed NMR shielding values have been averaged according to the weights of the Boltzmann analysis. Because calculated shielding constants have been referenced to chemical shifts obtained from ^1H - ^{13}C HSQC experiments, the MOSS-DFT database only contains ^{13}C - ^1H pairs of covalently bonded atoms. The calculated shielding

constants of methyl and methylene protons have been averaged to reduce misassignments in the case of magnetically inequivalent protons. In our calculations, about 4% of the atoms had chemical shift deviations larger than 7 ppm and 0.6 ppm for ^{13}C and ^1H , respectively. A closer inspection of the problematic compounds suggests in some cases misassignments in the experimental spectra or obvious problems in conformational sampling, such as errors in dihedral angles, especially for some tertiary carbon atoms. However, for the majority of outliers it was not easily possible to identify the specific cause for the deviation, such as less common functional groups. To prevent biased linear regression parameters, the molecules containing those atoms have been removed from the database, and the linear fit was repeated for the remaining 176 molecules. Finally, to convert the shielding constants to chemical shifts, a motif-specific linear regression approach was used according to

$$\delta = a_i \sigma_{\text{iso}} + b_i \quad (2)$$

where σ_{iso} denotes the shielding constant; a_i and b_i are model parameters for motif i (details on motifs are given below) and δ is the chemical shift value.

3. Results and Discussion

3.1 Motif-Specific Linear Regressions for the More Accurate Chemical Shift Prediction

Table 1 shows the definition and names of motifs that are used to convert the calculated shielding constants to chemical shifts. We assigned each ^{13}C - ^1H pair to one of the motifs based on the chemical environment of its carbon atom. That is, we determined in a first step whether the carbon atom is sp^3 or sp^2 hybridized, and subsequently, whether it is bonded to any element other than carbon or hydrogen. For sp^2 carbons, we further discriminated between aromatic and aliphatic carbons, where we assigned aromatic atoms to ar_{het} and $\text{ar}_{\text{nonhet}}$ depending on whether they are bonded to a hetero atom or not. Aliphatic sp^2 atoms were assigned to the sp^2 motif. Due to the relatively small number of atoms within this motif, it was not further divided into hetero or non-hetero bonded carbons. Note, these definitions are mutually exclusive, that is, every ^{13}C - ^1H pair is assigned to only one of the motifs. Our motivation for the present classification is threefold: we aim (I) to correct for errors due to the implicit modeling of solute-water interactions that will affect carbons bonded to heteroatoms differently than carbon atoms bonded to other carbon atoms, (II) to identify and obtain groups of atoms with superior or inferior prediction accuracy, and optimize their scaling parameters to further reduce errors within these groups, and (III) to account for slightly different slopes of aliphatic and aromatic carbons and protons. As part of the fitting process, we carried out a quadratic fit for each of the motifs and found no significant improvements.

The RMSD values and slopes of the linear regressions for the best performing method, B97-2/pcS-3, are reported in Table 2, where the considered RMSD intervals were limited to 0–7 ppm (^{13}C) and 0–0.6 ppm (^1H), respectively. Our findings show a good overall performance of the MOSS-DFT approach as indicated by RMSD values of 1.93 and 0.154 ppm for ^{13}C and ^1H , respectively. A particularly good prediction is achieved for motif $\text{sp}^3_{\text{nonhet,met}}$ with

RMSD values of 1.15 and 0.079 for ^{13}C and ^1H , respectively. Therefore, the chemical shift prediction of methyls that are not bonded to any heteroatom should be given a higher weight when such information is used for the identification and validation of unknown metabolites.

As a general trend, a higher prediction accuracy is achieved for ^{13}C - ^1H pairs that are not directly bonded to a heteroatom, with carbons in the group ar_{het} being a notable exception. The worst prediction results are found for sp^2 carbons and protons, as well as for protons within the ar_{het} group. Nevertheless, with the MOSS-DFT prediction approach we obtained an accuracy that is comparable to earlier studies of organic solvents using global scaling factors and intercepts.^{20, 24, 28–30} We also note that in our case the RMSD values for global scaling are 2.36 and 0.170 ppm for ^{13}C and ^1H , respectively, and thus significantly larger than those for the motif-specific scaling. To cross-validate our approach, a new training set was created by randomly excluding 20% of the compounds from the original MOSS-DFT training set and assigning them to a test set. The obtained linear regression parameters for the new training set and the RMSD values of the test set are reported in Table S3. We find overall RMSD values of 1.94 ppm and 0.133 ppm for ^{13}C and ^1H , respectively, which are well comparable to the ones reported above. In addition, a comparison of the linear regression parameters in Table S3 with the original MOSS-DFT parameters shows only marginal differences, which confirms that the MOSS-DFT parameters are robust, i.e. they are insensitive to the exact choice of the database. As a result, a good prediction accuracy can be expected when applying the MOSS-DFT parameters to targets outside of the training set, provided that the conformational ensemble of the query compound has been obtained in a similar way.

Figures 1 and 2 show the correlation between shielding constants and experimental chemical shifts within the considered RMSD interval. The high R^2 coefficients indicate very good correlations considering the narrow chemical shift ranges of most of the motifs. Usually, better correlations are found for ^{13}C than for ^1H . This is not unexpected, since the chemical shift ranges are much smaller in the latter case. Furthermore, there are significant differences between global and motif-specific fits. The largest discrepancies between global and motif-specific slopes are found for $\text{ar}_{\text{nonhet}}$ and ar_{het} , indicating larger systematic errors in these cases. It has been previously proposed for Hartree-Fock that the larger degree of electron correlation in aromatic compounds is responsible for larger systematic errors,⁴⁹ but this is still under debate for DFT. Nevertheless, Table 2 shows that the RMSD values of the aromatic groups are mostly lower than those for $\text{sp}^3_{\text{nonhet}}$, which illustrates the importance of the motif-specific corrections used here. Another interesting observation is found for motif $\text{sp}^3_{\text{nonhet,met}}$: the correlation plots for ^{13}C and ^1H show almost parallel ‘best fit’ lines for global and motif-specific regressions. At this point, it is unclear whether this result is intrinsic to our model (i.e. GIAO-DFT with implicit water model and the use of optimized geometries) or whether it is of purely statistical nature. The use of a larger sample size with a concomitant reduction of the spread of the error distribution might be able to resolve this issue.

3.2 Choice of Basis Set and Exchange-Correlation Functional

The computational cost of chemical shift predictions is especially crucial in the field of metabolomics as often large sets of candidate compounds exist for which predictions are needed. Clearly, the computational time required for any QC method to calculate shielding constants will depend on the size of the basis set. To find an acceptable tradeoff between computational cost and prediction accuracy, the dependence of RMSD values for ^{13}C and ^1H chemical shifts on the basis set size has been examined. The results for the sequence pcS-1, pcS-2 and pcS-3 are shown in Table 3. Note that the average computational time for a single NMR shielding calculation increases from 1 minute for pcS-1 to around 4 hours for pcS-3 on 12 cores. It is found that the RMSD value for ^{13}C chemical shifts decreases by 0.09 ppm when the pcS-2 basis set is used instead of pcS-1. The use of pcS-3 only yields a marginal improvement of 0.03 ppm. Although, on an absolute scale, the differences between RMSD values of ^1H are comparable to those of ^{13}C , it is clear that changes in the second digit after the decimal point are already a major accuracy gain for ^1H . Like for ^{13}C , our results indicate a consistent improvement for ^1H chemical shifts when going to larger basis sets, where the difference between pcS-1 and pcS-2 is as large as 0.058 ppm. However, unlike for ^{13}C , the use of B97-2/pcS-3 still leads to a small improvement by 0.01 ppm compared to B97-2/pcS-2. It is also interesting to note that there is no obvious relation between a scaling factor closer to one and a lower RMSD value (see Tables S1 and S2). Nevertheless, one observes convergence of the scaling factors in the case of ^{13}C for pcS-2 to pcS-3. Thus, we can conclude that the accuracy gain results from an improved correlation between calculated shielding constants and experimental chemical shifts, but not from the reduction of the method's systematic error as reflected by the scaling factor.

In addition to the basis set convergence, we compared the performances of two other exchange-correlation functionals. The results are reported in Table 3. The first functional, B3LYP, is one of the most widely used hybrid DFT functionals for organic molecules, whereas the second, BLYP, is a GGA functional often used in explicit solvent *ab-initio* molecular dynamics simulations and NMR shielding calculations due to its numerical efficiency. Compared to B97-2/pcS-2, both functionals perform slightly worse for ^{13}C with RMSDs of 0.07 and 0.66 ppm for B3LYP and BLYP, respectively. Interestingly, the prediction performance of both functionals for ^1H is consistent with that of B97-2. Unlike the basis set dependence, our findings show a correlation between a lower RMSD and a slope closer to one (see Tables S1 and S2), which indicates a smaller systematic error.

3.3 Comparison with Empirical Predictions

Since the time needed for QC predictions is typically orders of magnitudes larger than for empirical predictions, the former need to deliver a substantial accuracy increase to justify their use in metabolomics studies. To verify this, we compared the RMSD values of our MOSS-DFT method to empirical predictions in Table 4. Likewise, for the QC results above, all molecules that contained $^{13}\text{C}/^1\text{H}$ atoms, whose empirically predicted shifts deviated by more than 7 ppm / 0.6 ppm, have been omitted in the RMSD calculation. We note, however, that a completely unbiased comparison between both methods is not possible in our case, because it would require an independent validation set and therefore the knowledge of the training set of the NMRPredict program, which is not available. Our findings show an

improvement of more than 1.2 ppm for ^{13}C and 0.1 ppm for ^1H of the QC method in terms of RMSD over the empirical NMRPredict program. Even by using only a global correction, our QC predictor still outperforms the empirical predictor. However, predictions that are closely related to those in the training set are usually predicted by the empirical method with a very high accuracy, whereas structures that are dissimilar to the training set tend to be predicted significantly less accurately. We also investigated the correlation between the predictions by MOSS-DFT and NMRPredict (Figure 3) and find that there is no obvious correlation between the two. Nevertheless, both methods seem to be complementary in the sense that there are only few predictions that have large errors in both cases. The low probability of large errors for either method could be used as additional scoring information as part of the identification protocol. For example, if both predictors show a relatively large error for a candidate structure, it can be assigned a lower matching score.

4. Conclusion

In this work, we have presented a new DFT-based chemical shift prediction approach in aqueous solution, MOSS-DFT, that is specifically suited for the analysis of metabolites. To correct for systematic errors, a motif-specific model was introduced, which goes beyond a common global scaling correction. Our approach is based on a set of 176 molecules, where ^{13}C - ^1H atom pairs have been classified into 6 motifs based on hybridization, aromaticity, and heteroatom bonding of the carbon atom. Subsequently, linear regression parameters have been derived separately for each of the motifs. We found a total RMSD value with respect to experimental data of 1.93/0.154 ppm for $^{13}\text{C}/^1\text{H}$. This value is 0.43/0.016 ppm lower than that of the global linear correction. As a general trend, predictions of atoms that are not bonded to a heteroatom exhibit a lower error. The best motif-specific RMSD values were obtained for $\text{sp}^3_{\text{nonhet,met}}$ and $\text{ar}_{\text{nonhet}}$ with 1.15/0.079 ppm and 1.31/0.118 ppm for $^{13}\text{C}/^1\text{H}$ atoms. The non-aromatic sp^2 atoms have the highest RMSD (3.03 ppm) for ^{13}C , whereas for ^1H the highest RMSD of 0.239 ppm was found for ar_{het} . A comparison of different functional/basis set combinations suggests that B97-2/pcS-2 is most economical, although B97-2/pcS-3 is slightly more accurate for ^1H . In the case of ^1H , no significant improvement is found for B97-2 and B3LYP functionals over BLYP. A limitation of the MOSS-DFT approach is for molecules that form intramolecular hydrogen bonds. Therefore, future developments and applications of numerically efficient protocols should focus on the improved sampling and selection of relevant conformations for further improvements toward accurate calculations of chemical shifts of a growing spectrum of synthetic and naturally occurring molecules.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

F.H. thanks the Fonds der Chemischen Industrie for a Kekulé fellowship. This work was supported by the National Institutes of Health (grant R01GM066041 to R.B.). We thank the Ohio Supercomputer Center and the Martin-Luther University Halle-Wittenberg for providing us high performance computing resources.

References

1. Nagana Gowda GA, Raftery D. Can NMR solve some significant challenges in metabolomics? *J Magn Reson.* 2015; 260:144–60. [PubMed: 26476597]
2. Fan TW, Lane AN. Applications of NMR spectroscopy to systems biochemistry. *Prog Nucl Magn Reson Spectrosc.* 2016; 92–93:18–53.
3. Markley JL, Brüschweiler R, Edison AS, Eghbalnia HR, Powers R, Raftery D, Wishart DS. The future of NMR-based metabolomics. *Curr Opin Biotechnol.* 2017; 43:34–40. [PubMed: 27580257]
4. Bingol K, Bruschiweiler-Li L, Li D, Zhang B, Xie M, Brüschweiler R. Emerging new strategies for successful metabolite identification in metabolomics. *Bioanalysis.* 2016; 8:557–73. [PubMed: 26915807]
5. Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev.* 2011; 40:387–426. [PubMed: 20717559]
6. Beckonert O, Keun HC, Ebbels TM, Bundy J, Holmes E, Lindon JC, Nicholson JK. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc.* 2007; 2:2692–703. [PubMed: 18007604]
7. Kim HK, Choi YH, Verpoorte R. NMR-based metabolomic analysis of plants. *Nat Protoc.* 2010; 5:536–49. [PubMed: 20203669]
8. Palmnas MS, Vogel HJ. The future of NMR metabolomics in cancer therapy: towards personalizing treatment and developing targeted drugs? *Metabolites.* 2013; 3:373–96. [PubMed: 24957997]
9. Bingol K, Li DW, Bruschiweiler-Li L, Cabrera OA, Megraw T, Zhang F, Brüschweiler R. Unified and isomer-specific NMR metabolomics database for the accurate analysis of (13)C-(1)H HSQC spectra. *ACS Chem Biol.* 2015; 10:452–9. [PubMed: 25333826]
10. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, et al. HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res.* 2013; 41:D801–7. [PubMed: 23161693]
11. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, et al. BioMagResBank. *Nucleic Acids Res.* 2008; 36:D402–8. [PubMed: 17984079]
12. NMRPredict. Modgraph Consultants Ltd;
13. Mnova. Mestrelab Research;
14. ACD/NMR Predictors. Advanced Chemistry Development;
15. Kwan EE, Liu RY. Enhancing NMR prediction for organic compounds using molecular dynamics. *J Chem Theory Comput.* 2015; 11:5083–5089. [PubMed: 26574306]
16. Dracinsky M, Moller HM, Exner TE. Conformational sampling by ab initio molecular dynamics simulations improves NMR chemical shift predictions. *J Chem Theory Comput.* 2013; 9:3806–15. [PubMed: 26584127]
17. Bekcioglu-Neff G, Allolio C, Desmukh YS, Hansen MR, Sebastiani D. Dynamical dimension to the Hofmeister series: insights from first-principles simulations. *ChemPhysChem.* 2016; 17:1166–73. [PubMed: 26864593]
18. Elgabarty H, Schmieder P, Sebastiani D. Unraveling the existence of dynamic water channels in light-harvesting proteins: alpha-C-phycoerythrin in vitro. *Chem Sci.* 2013; 4:755–763.
19. Banyai DR, Murakhtina T, Sebastiani D. NMR chemical shifts as a tool to analyze first principles molecular dynamics simulations in condensed phases: the case of liquid water. *Magn Reson Chem.* 2010; 48(Suppl 1):S56–60. [PubMed: 21104763]
20. Lodewyk MW, Siebert MR, Tantillo DJ. Computational prediction of 1H and 13C chemical shifts: a useful tool for natural product, mechanistic, and synthetic organic chemistry. *Chem Rev.* 2012; 112:1839–62. [PubMed: 22091891]
21. Willoughby PH, Jansma MJ, Hoye TR. A guide to small-molecule structure assignment through computation of 1H and 13C NMR chemical shifts. *Nat Protoc.* 2014; 9:643–60. [PubMed: 24556787]
22. Tantillo DJ. Walking in the woods with quantum chemistry--applications of quantum chemical calculations in natural products research. *Nat Prod Rep.* 2013; 30:1079–86. [PubMed: 23793561]

23. Cormanich RA, Buhl M, Rittner R. Understanding the conformational behaviour of Ac-Ala-NHMe in different media. A joint NMR and DFT study. *Org Biomol Chem*. 2015; 13:9206–13. [PubMed: 26219244]
24. Flaig D, Maurer M, Hanni M, Braunger K, Kick L, Thubauville M, Ochsenfeld C. Benchmarking hydrogen and carbon NMR chemical shifts at HF, DFT, and MP2 Levels. *J Chem Theory Comput*. 2014; 10:572–8. [PubMed: 26580033]
25. Gauss J, Werner HJ. NMR chemical shift calculations within local correlation methods: the GIAO-LMP2 approach. *Phys Chem Chem Phys*. 2000; 2:2083–2090.
26. Loibl S, Schutz M. NMR shielding tensors for density fitted local second-order Moller-Plesset perturbation theory using gauge including atomic orbitals. *J Chem Phys*. 2012; 137:084107. [PubMed: 22938218]
27. Maurer M, Ochsenfeld C. A linear- and sublinear-scaling method for calculating NMR shieldings in atomic orbital-based second-order Moller-Plesset perturbation theory. *J Chem Phys*. 2013; 138:174104. [PubMed: 23656111]
28. Konstantinov IA, Broadbelt LJ. Regression formulas for density functional theory calculated ¹H and ¹³C NMR chemical shifts in toluene-d₈. *J Phys Chem A*. 2011; 115:12364–72. [PubMed: 21966955]
29. Pierens GK. ¹H and ¹³C NMR scaling factors for the calculation of chemical shifts in commonly used solvents using density functional theory. *J Comput Chem*. 2014; 35:1388–94. [PubMed: 24854878]
30. Benassi E. Benchmarking of density functionals for a soft but accurate prediction and assignment of ¹H and ¹³C NMR chemical shifts in organic and biological molecules. *J Comput Chem*. 2017; 38:87–92. [PubMed: 27796077]
31. Sarotti AM, Pellegrinet SC. A multi-standard approach for GIAO ¹³C NMR calculations. *J Org Chem*. 2009; 74:7254–7260. [PubMed: 19725561]
32. Zhu T, Zhang JZ, He X. Automated fragmentation QM/MM calculation of amide proton chemical shifts in proteins with explicit solvent model. *J Chem Theory Comput*. 2013; 9:2104–14. [PubMed: 26583557]
33. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform*. 2011; 3:33. [PubMed: 21982300]
34. MacroModel. Schrödinger Release 2014. Schrödinger; New York: 2014.
35. Banks JL, Beard HS, Cao Y, Cho AE, Damm W, Farid R, Felts AK, Halgren TA, Mainz DT, Maple JR, et al. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J Comput Chem*. 2005; 26:1752–80. [PubMed: 16211539]
36. Chang G, Guida WC, Still WC. An internal coordinate Monte-Carlo method for searching conformational space. *J Am Chem Soc*. 1989; 111:4379–4386.
37. Saunders M, Houk KN, Wu YD, Still WC, Lipton M, Chang G, Guida WC. Conformations of cycloheptadecane - a comparison of methods for conformational searching. *J Am Chem Soc*. 1990; 112:1419–1427.
38. Frisch, MJ., Trucks, GW., Schlegel, HB., Scuseria, GE., Robb, MA., Cheeseman, JR., Scalmani, G., Barone, V., Mennucci, B., Petersson, GA., et al. Gaussian 09. Gaussian, Inc; Wallingford, CT, USA: 2009.
39. Becke AD. Density-functional thermochemistry. III. The role of exact exchange. *J Chem Phys*. 1993; 98:5648–5652.
40. Grimme S, Antony J, Ehrlich S, Krieg H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys*. 2010; 132:154104. [PubMed: 20423165]
41. Weigend F, Ahlrichs R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. *Phys Chem Chem Phys*. 2005; 7:3297–305. [PubMed: 16240044]
42. Barone V, Cossi M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J Phys Chem A*. 1998; 102:1995–2001.

43. Cossi M, Rega N, Scalmani G, Barone V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J Comput Chem.* 2003; 24:669–81. [PubMed: 12666158]
44. London F. Quantum theory of interatomic currents in aromatic compounds. Théorie quantique des courants interatomiques dans les combinaisons aromatiques. *J Phys Rad.* 1937; 8:397–409.
45. McWeeny R. Perturbation theory for the Fock-Dirac density matrix. *Phys Rev.* 1962; 126:1028.
46. Ditchfield R. Self-consistent perturbation theory of diamagnetism: I. A gauge-invariant LCAO method for NMR chemical shifts. *Mol Phys.* 1974; 27:789–807.
47. Wolinski K, Hinton JF, Pulay P. Efficient implementation of the gauge-independent atomic orbital method for NMR chemical-shift calculations. *J Am Chem Soc.* 1990; 112:8251–8260.
48. Cheeseman JR, Trucks GW, Keith TA, Frisch MJ. A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J Chem Phys.* 1996; 104:5497–5509.
49. Facelli, J. *Encyclopedia of Nuclear Magnetic Resonance.* Grant, DM., Harris, RK., editors. Vol. 9. London: John Wiley & Sons; 2002. p. 323-333.
50. Wilson PJ, Bradley TJ, Tozer DJ. Hybrid exchange-correlation functional determined from thermochemical data and ab initio potentials. *J Chem Phys.* 2001; 115:9233–9242.
51. Jensen F. Basis set convergence of nuclear magnetic shielding constants calculated by density functional methods. *J Chem Theory Comput.* 2008; 4:719–27. [PubMed: 26621087]
52. Becke AD. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys Rev A.* 1988; 38:3098–3100.
53. Lee CT, Yang WT, Parr RG. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron-density. *Phys Rev B.* 1988; 37:785–789.
54. Miehlich B, Savin A, Stoll H, Preuss H. Results obtained with the correlation-energy density functionals of Becke and Lee, Yang and Parr. *Chem Phys Lett.* 1989; 157:200–206.

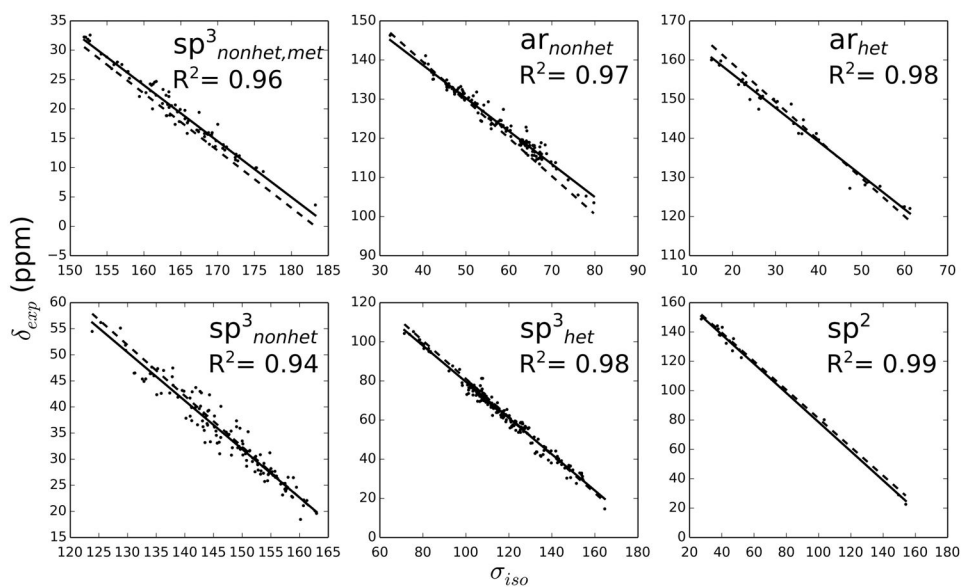


Figure 1. Experimental chemical shifts *versus* computed isotropic shielding constants of ^{13}C using the B97-2/pcS-3 method. The solid lines belong to the best fit for each motif, while the dashed lines correspond to the global fit.

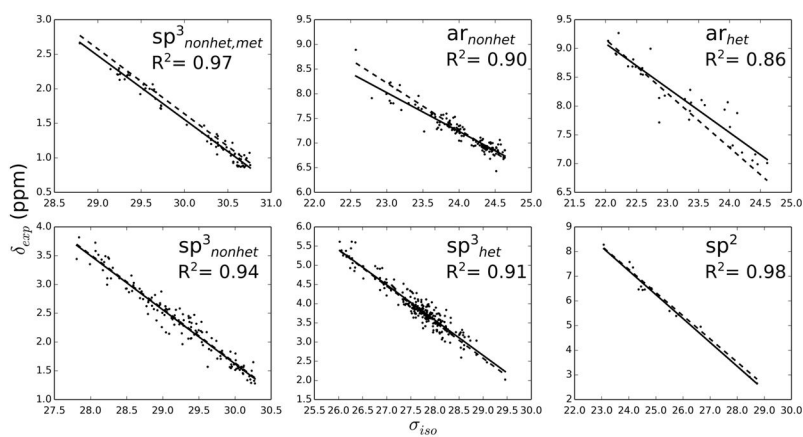


Figure 2. Experimental chemical shifts *versus* computed isotropic shielding constants of ^1H using the B97-2/pcS-3 method. The solid lines belong to the best fit for each type, while the dashed lines correspond to the global fit.

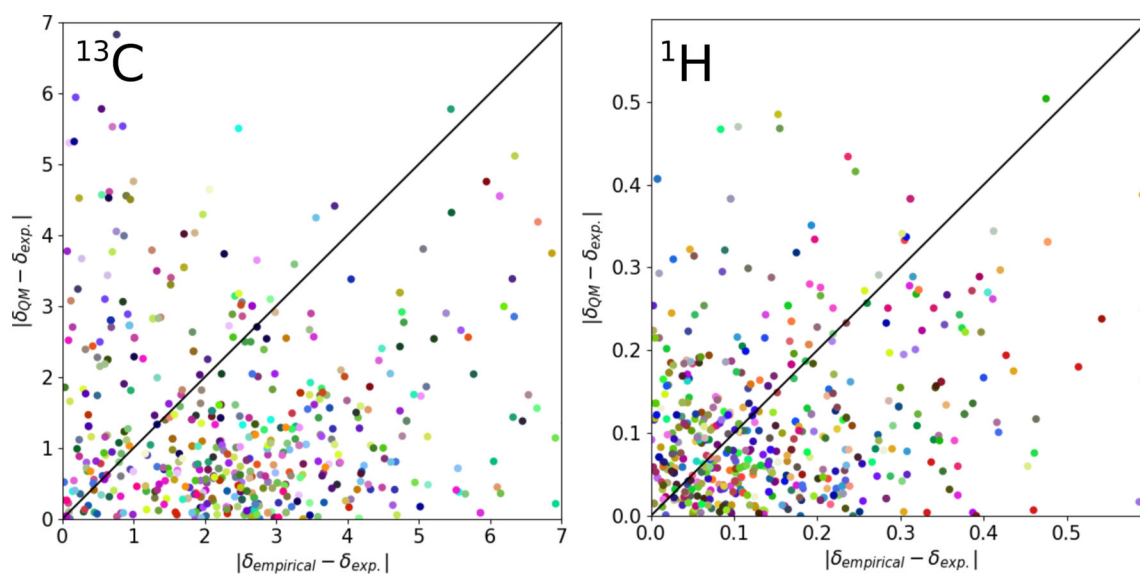


Figure 3.

Absolute deviations of predictions by MOSS-DFT *versus* NMRPredict for ^{13}C and ^1H . The diagonal is shown to guide the eye. Points marked with the same color belong to the same molecule. Note that the chemical shift error ranges for ^{13}C and ^1H have been limited to the intervals indicated in the figure.

Table 1

Definitions, Names, and Examples of the Molecular Motifs Used in the Linear Regression Approach.

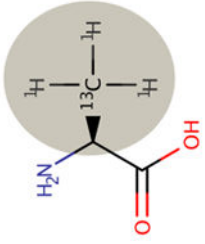
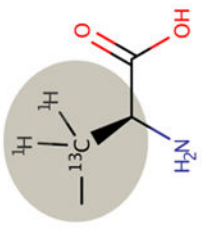
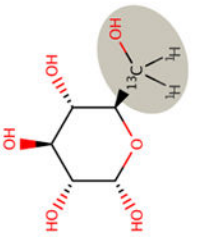
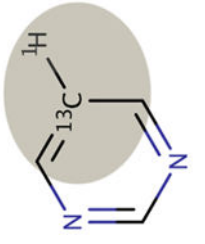
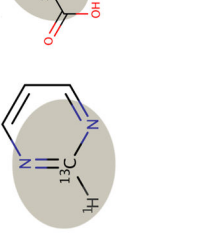
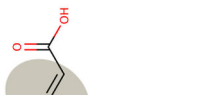

Motif Definition:	Methyl groups not bonded to a heteroatom	sp^3 carbon centers not bonded to a heteroatom, excluding methyl groups	sp^3 carbon centers bonded to a heteroatom	Aromatic carbon centers not bonded to a heteroatom	Aromatic carbon centers bonded to a heteroatom	sp^2 carbon centers	
Motif Name:	$sp^3_{nonhet, met}$	sp^3_{nonhet}	sp^3_{het}	ar_{nonhet}	ar_{het}	sp^2	
Examples:							

Table 2

RMSDs with respect to Experimental NMR Chemical Shifts along with Linear Regression Parameters for the B97-2/pcS-3 Method.

Nucleus	$sp^3_{nonhet,met}$	ar_{nonhet}	ar_{het}	sp^3_{nonhet}	sp^3_{het}	sp^2	total
a	-0.9532	-0.8433	-0.8634	-0.9277	-0.9261	-0.9936	
b	176.5288	172.4185	173.6563	171.0732	172.1137	177.9458	
RMSD	1.15	1.31	1.76	2.05	2.32	3.03	1.93
<hr/>							
a	-0.9244	-0.7854	-0.7725	-0.9399	-0.9181	-0.9677	
b	29.2860	26.0863	26.0812	29.8153	29.2769	30.4447	
RMSD	0.079	0.118	0.239	0.146	0.177	0.214	0.154
<hr/>							
No. of ^{13}C - 1H pairs	87	192	35	141	285	16	756

Table 3

RMS Deviations of Chemical Shifts for Different Exchange-Correlation Functionals and Basis Sets.

XC Functional / Basis Set	RMSD ¹³ C (ppm)	RMSD ¹ H (ppm)
B97-2 ⁵⁰ / pcS-1 ⁵¹	2.05	0.222
B97-2 / pcS-2 ⁵¹	1.96	0.164
B97-2 / pcS-3 ⁵¹	1.93	0.154
B3LYP ³⁹ / pcS-2	2.03	0.165
BLYP ⁵²⁻⁵⁴ / pcS-2	2.62	0.166

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

RMSD Values for MOSS-DFT (B97-2 / pcS-3) and NMRPredict Chemical Shift Predictions with respect to experiment.

¹³ C	RMSD _{MOSS-DFT}	1.93 ppm
	RMSD _{NMRPredict}	3.15 ppm*
¹ H	RMSD _{MOSS-DFT}	0.154 ppm
	RMSD _{NMRPredict}	0.255 ppm*

* Outliers above 7 ppm (¹³C) / 0.6 ppm (¹H) have been excluded

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript