PCCP

Accurate and Cost-Effective NMR Chemical Shift Predictions for Proteins using Molecules-in-Molecules Fragmentation-Based Method

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-09-2020-005064.R1
Article Type:	Paper
Date Submitted by the Author:	13-Nov-2020
Complete List of Authors:	Chandy, Sruthy; Indiana University Bloomington, Chemistry Thapa, Bishnu; Indiana University Bloomington, Chemistry Raghavachari, Krishnan; Indiana University, Chemistry

Accurate and Cost-Effective NMR Chemical Shift Predictions for Proteins using Molecules-in-Molecules Fragmentation-Based Method

*Sruthy Kettidathil Chandy, Bishnu Thapa, and Krishnan Raghavachari**

Department of Chemistry, Indiana University, Bloomington, Indiana, U.S.A.

Abstract

We have developed an efficient protocol using our two-layer Molecules-in-Molecules (MIM2) fragment-based quantum chemical method for the prediction of NMR chemical shifts of large biomolecules. To investigate the performance of our fragmentation approach and demonstrate its applicability, MIM-NMR calculations are first calibrated on a test set of six proteins. The MIM2-NMR method yields a mean absolute deviation (MAD) from unfragmented full molecule calculations of 0.01 ppm for ^{1}H and 0.06 ppm for ^{13}C chemical shifts. Thus, the errors from fragmentation are only about 3% of our target accuracy of ~0.3 ppm for ^{1}H and 2-3 ppm for ^{13}C chemical shifts. To compare with experimental chemical shifts, a standard protocol is first derived using two smaller proteins 2LHY (176 atoms) and 2LI1 (146 atoms) for obtaining an appropriate protein structure for NMR chemical shift calculations. The effect of the solvent environment on the calculated NMR chemical shifts is incorporated through explicit, implicit, or implicit-explicit solvation models. The expensive first solvation shell calculations are replaced by a micro-solvation model in which only the immediate interaction between the protein and the explicit solvation environment is considered. A single explicit water molecule for each amine and amide protons is found to be sufficient to yield accurate results for ^{1}H chemical shifts. The ^{1}H and ^{13}C NMR chemical shifts calculated using our protocol give excellent agreement with experiments for two larger proteins 2MC5 (the helical part with 265 atoms) and 3UMK (33 residue slice with 547 atoms). Overall, our target accuracy of ~0.3 ppm for ^{1}H and ~2-3 ppm for ^{13}C has been achieved for the larger proteins. The proposed MIM-NMR method is accurate and computationally cost-effective and should be applicable to study a wide range of large proteins.

1. Introduction

Analytical techniques such as NMR, X-ray crystallography, and electron microscopy are widely used to determine the structures of biomolecules.[1-5] Among these techniques, nuclear magnetic resonance (NMR) stands out as a particularly versatile and powerful spectroscopic tool for understanding the structure and dynamics of biomolecules for applications in scientific research, medicine, and industry.[6-11] Often, NMR experiments on proteins are conducted in solution at the physiological pH and can help us determine the functions of proteins in biological systems under realistic conditions. While NMR is widely used to deduce the structures and properties of proteins, the data obtained, however, are often noisy, and hence, there is significant uncertainty about the actual protein structure.[12] In such cases, computational methods present a reliable alternative to get accurate structures of proteins by utilizing the available experimental NMR results in conjunction with quantum chemical techniques.[13-16]

The primary data in NMR experiments include chemical shifts, coupling constants, and relative integrated intensities. Among these, the chemical shift includes information about the local magnetic as well as chemical environmental effects, and can be effectively studied using quantum chemical computational tools, including semi-empirical, *ab initio*, and density functional theory (DFT) methods.[14, 17-31] However, empirical methods, such as SHIFTS, SHIFTX2, CAMSHIFT, and PROSHIFT, that use parameters derived from fitting empirical formulae to the known experimental chemical shifts, are most widely used to determine the structures of larger organic molecules and biomolecules.[32-35] While such methods have been shown to perform well for systems similar to those used in the parameterization (e.g., standard amino acids in their native environment), their performance may have limitations for nonstandard systems such as mutated residue side chains, metal cofactors, and protein inhibitors.[36]

As an alternative, first principles-based electronic structure calculations can, in principle, provide accurate chemical shifts, comparable to experimentally determined values, independent of the chemical composition of the molecule.[37-39] However, the reliability of the NMR results depends on the accuracy of the quantum mechanical (QM) method used and proper modeling of the solvent environment.[40] For the larger biomolecules such as proteins that contain several hundreds or thousands of atoms, the computational cost of the highly accurate QM methods becomes intractable. Therefore, as is often the case, there is a trade-off between accuracy and computational

cost while dealing with large molecules. For example, accurate *ab initio* methods such as MP2 or CCSD, used in conjunction with gauge-independent atomic orbital (GIAO) NMR calculations, are affordable only for chemical shifts of small molecules, whereas DFT methods with reasonable accuracy but much more cost-effectiveness, are often used for larger biomolecular systems.[41-46]

In addition to the accuracy of the QM method used, incorporating the solvation effects can play a crucial role in obtaining reliable results compared to the experimental NMR chemical shifts.[47] Commonly, solvation effects are incorporated as a polarizable dielectric continuum, instead of including explicit solvent molecules, to lower the computational cost. However, such models neglect the non-bonded interactions (e.g., hydrogen bonding) between the solute and solvent molecules, which can be essential for predicting accurate chemical shifts in NMR spectroscopy. Indeed, explicit solvation is shown to provide significant improvements in the chemical shifts of small molecular species in recent works.[48, 49] In such a scenario, a full explicit solvation model, or an implicit model containing a few nearby solvent molecules, would be necessary.

Although numerous studies have been reported in the literature, many of the existing empirical and quantum chemical methods still fail to accurately interpret NMR spectra in the case of large biomolecules with more than one thousand atoms.[36] In particular, since several secondary interactions are possible with overlapping spectral features, the correct interpretation of the NMR spectra of such large biomolecules can be quite difficult. For example, Sumowski *et al.* found that the QM methods are more sensitive to electronic and structural changes when compared to existing empirical methods.[50] The major limitation of the current QM-based approaches is that such protocols become too expensive as the system size becomes larger.[51] Since full quantum chemical computations for large proteins are currently not feasible, most of the previous studies have been carried out only on localized truncated structural models to obtain the NMR chemical shifts.[52, 53] Recently, fragmentation-based hybrid methods are evolving as highly efficient tools for linear-scaling QM calculations of large systems.[54, 55] Larger molecules are fragmented into smaller pieces, and by employing QM calculations, the wave-function, energy, and other energy derivatives (i.e., molecular properties) of each fragment are calculated. Then the results of the fragments are combined to extrapolate to the results for the full molecule.[56-68] Fragmentation methods rely on the chemical locality of macromolecular systems, assuming the local region of a

macromolecule is only slightly affected by the atoms that are far away from the region of interest.[36, 52, 60, 69] The earlier work by Scheurer *et al*. used DFT calculations on manually generated fragments to calculate the anisotropy tensors for chemical shielding.[70] Subsequently, the local nature of nuclear shielding tensors has been used with a QM/MM framework by Cui *et al*.[53] Further, adjustable density matrix assembler (ADMA) fragmentation-based method, fragment molecular orbital (FMO) method, combined fragmentation method (CFM), generalized energy-based fragmentation(GEBF), systematic molecular fragmentation analysis (SMFA), automated fragmentation quantum mechanics/molecular mechanics approach (AF-QM/MM) and fragment based electronic structure approach have been developed by different groups to compute the NMR chemical shifts of proteins and nucleic acids.[38, 71-78] Recent work from the Beran group demonstrates the effect of solvation in prediction of NMR shielding tensors for molecular crystals using PCM-embedded fragmentation approach.[79]

In this study, we have used our multilayer Molecules-in-Molecules (MIM) fragmentation-based method, which shares a similar working principle with the popular ONIOM approach, to calculate the NMR spectra of selected illustrative polypeptides.[80-83] Our multilayer MIM scheme also offers significant flexibility in choosing the combinations of levels of theory for calculating the desired molecular property to lower the computational cost substantially. MIM method has previously shown excellent performance on a range of spectroscopic studies including infrared (IR), Raman, vibrational circular dichroism (VCD), and Raman optical activity spectra on large systems.[84-86] As discussed above, since the nuclear shielding is a local property, applying high-level QM methods on smaller fragment sizes to include the most important components, and capturing the long-range interactions through efficient low level theory calculations, makes MIM an accurate and cost-effective method to predict the NMR chemical shifts of large biomolecules. In this study, we have expanded our previously developed MIM-NMR method[40] by carefully calibrating the combinations of various levels of theory for the precise evaluation of NMR chemical shifts. Furthermore, we also evaluate the effect of including conformational changes as well as the effect of structural minimization on the computed NMR spectra. Additionally, we present an accurate and cost-effect approach of incorporating solvation effects on the calculated NMR chemical shifts.

2. Methods

2.1 Molecules-in-molecules (MIM) method

All MIM and MIM-NMR calculations were performed using an external perl module and the Gaussian16 program suite.[87] The details about the working principles of our MIM fragment-based approach, different fragmentation schemes, and capabilities of our method have been described in previous publications.[40, 83-86, 88-90] Therefore, only a brief and relevant discussion will be given here. In MIM, initial non-overlapping fragments, called "monomers", are formed by cutting single bonds between heavy (non-hydrogen) atoms. In the case of proteins, we keep the peptide C–N bonds intact due to their partial double bond character. In this work, we have employed a fragmentation scheme where we only cut the C–Cα bond and keep the sidechain and peptide backbone together to form non-overlapping monomers. Neighboring monomers are combined to form primary and derivative subsystems (*vide infra*) to capture the interactions between the monomers.

Throughout this study, we have employed a two-layer MIM approach (MIM2). In MIM2, two fragmentation parameters and two levels of theory are used to compute the relevant properties of the molecule. The primary subsystems formed with a small fragmentation parameter (*r*) are calculated with both the high and low levels of theory, and those with a large parameter (*R*, full molecule in this study) accounting for long-range interactions are calculated only at a low level of theory. With the smaller fragmentation parameter (*r*), the primary subsystems are formed by combining four of the adjacent monomers resulting in a tetramer (or tetrapeptide) subsystem. These tetrapeptide primary subsystems are ideal for the NMR calculations since their size is small enough to perform the NMR calculations at the high-level of theory without being a computational bottleneck while capturing some of the *intramolecular hydrogen bonding* interactions. Since the primary subsystems are formed by starting from each of the monomers, there are overlapping parts that need to be accounted for. To account for the over-counting of the overlapping parts, derivative subsystems are formed using the *inclusion-exclusion principle*. All the remaining missing inter-subsystem interactions are captured at a lower level of theory. The truncated bonds in the subsystems are saturated with link-hydrogen atoms. MIM2 energy can be written, similar to the standard ONIOM extrapolation expression, as shown in equation 1.

$$E^{MIM2} = E^{r}_{high} - E^{r}_{low} + E^{R}_{low} \; (r << R) \qquad (1)$$

Here (*r*) and (*R*) represent generalizations of the "model system" and "real system" as in the standard ONIOM calculations. Thus, E^r_{high}, E^r_{low}, E^R_{low}, represent the generalized E_{mh}, E_{ml}, E_{rl} in the ONIOM energy expression. As has been described previously, the energy summation for the high and low levels of theory is carried out according to the *inclusion-exclusion principle*, taking into account the appropriate signs of the energy terms involving the different primary and derivative subsystems. [40, 83-86] For example,

$$E_{low} = \sum_i E_l^i - \sum_{i<j} \left| E_l^i \cap E_l^j \right| + \sum_{i<j<k} \left| E_l^i \cap E_l^j \cap E_l^k \right| - \cdots + (-1)^{n-1} \left| E_l^1 \cap \ldots \cap E_l^n \right| \quad (2)$$

*w*here E_l^i represents the energy of the *i*th fragment at the low level of theory.

For the initial calibration calculations, MIM2[*m*PW1PW91/6-311G(d,p):*m*PW1PW91/6-31G(d)] method is used with *m*PW1PW91/6-311G(d,p) in high layer and *m*PW1PW91/6-31G(d) in low layer. Four different combinations of DFT methods [(i) *m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G, (ii) CAM-B3LYP-D3BJ/6-311++G(2d,2p):CAM-B3LYP-D3BJ/6-31G, (iii) B3LYP-D3BJ/6-311++G(2d,2p):B3LYP-D3BJ/6-31G, and (iv) *ω*B97X-D/6-311++G(2d,2p):*ω*B97X-D/6-31G] are used for the MIM2-NMR calculations for a test set of proteins to develop the protocol and to apply to a larger test set for validation. Note that we have used four popular density functionals and standard Pople-style basis sets for all our assessments in this work. Among these functionals, *m*PW1PW91 has previously been shown to be quite accurate for the calculation of NMR chemical shifts. We also note that empirical dispersion corrections (as in Grimme's D3 corrections) do not contribute to the NMR shielding tensor, since the correction is only a function of nuclear positions and not the external magnetic field or the nuclear magnetic moments. In addition, we note that some special purpose basis sets have been developed for the calculation of NMR chemical shifts.[91-93] However, since the main idea of our paper is to showcase the performance of the MIM method in predicting NMR chemical shifts using an efficient micro-solvation model, we have used standard basis sets and did not explore much on the effect of different basis sets on the calculated NMR chemical shifts. However, we do expect that the basis set error should be independent of the MIM protocol that we develop in this paper.

2.2 NMR calculations

For the NMR-GIAO method, isotropic shielding tensor, σ^N for atom *N*, is given as the second derivative of the electronic energy E, with respect to the external magnetic field B, and the nuclear magnetic moment m_N.

$$\sigma_{ij}^N = \left[\frac{\partial^2 E}{\partial B_i \partial m_{N_j}}\right]_{B=0} \quad (3)$$

σ_{ij}^N is the ij^{th} component of the shielding tensor, B_i is the i^{th} component of the external magnetic field and m_{N_j} is the j^{th} component of magnetic moment of the nucleus *N*.

In MIM2, Isotropic shielding tensor for all the atoms are calculated using a general expression,

$$\sigma_{ij}^N = \left[\frac{\partial^2 E_{total}}{\partial B_i \partial m_{N_j}}\right]_{B=0} = \frac{\partial^2 E_{rl}}{\partial B_i \partial m_{N_j}} - \frac{\partial^2 E_{ml}}{\partial B_i \partial m_{N_j}} + \frac{\partial^2 E_{mh}}{\partial B_i \partial m_{N_j}} \quad (4)$$

The atomic NMR shielding constant is one-third of the sum of the trace of the atomic shielding tensors from equation (3). $\sigma_{i,,}$ which is the isotropic chemical shift, is subtracted from the corresponding standard reference value (σ_{ref}), to yield the chemical shift of each atomic species. For ^{1}H and ^{13}C, the chemical shift is calculated using tetramethylsilane (TMS) as the reference. For ^{15}N and ^{17}O, NH_3 and H_2O molecules, respectively, are taken as the references.

$$\delta_i = \sigma_{ref} - \sigma_i \quad (5)$$

The contribution of the different conformers to the total NMR chemical shift value is calculated according to their weights from a Boltzmann distribution. Thus, the percentage mole fraction, P_i, of the i^{th} conformer from the total number of conformers can be calculated using equation (6).

$$P_i = \frac{e^{\frac{-E_i}{kT}}}{\sum_j e^{\frac{-E_j}{kT}}} \quad (6)$$

2.3 Solvation models

The solvent environment for the MIM-NMR calculation is incorporated using implicit, and explicit-implicit solvent models, using equation (7). For implicit solvation, SMD-SCRF[94] implicit solvation model is used.

$$E_{Total}^{Implicit} = E_{rl}^{Implicit} - E_{ml}^{Implicit} + E_{mh}^{Implicit} \quad (7)$$

In our explicit-implicit solvation model, the explicit solvent molecules are included only in the high layer that is computed at the high level of theory including implicit solvation, while only the implicit solvation model is included in the calculations with the low level of theory, as shown as in equation (8). The assumption here is that the tetramer primary subsystems with explicit-implicit solvation model can accurately model the local intramolecular hydrogen bonding interactions as well as intermolecular explicit interactions with the solvent (*vide infra*) at the high level of theory. The missing long-range interactions in the high layer are captured using the low layer via the implicit solvation effects. As our results show, this micro-solvation model is an elegant approach to significantly lower the computational cost of performing a full molecule calculation with explicit solvation while maintaining a high accuracy. This approach is in line with some of the recent studies suggesting that the amide protons (H^N) and ^{15}N are highly sensitive to the solvation environment, and that a small number of directly hydrogen bonded explicit water molecules are sufficient to accurately determine local molecular properties in the aqueous medium.[71, 95-99]

$$E_{Total}^{Solvation} = E_{rl}^{Implicit} - E_{ml}^{Implicit} + E_{mh}^{Explicit-Implicit} \quad (8)$$

In our explicit-implicit solvation model, the short-range hydrogen-bonding interactions are captured by including *one explicit water molecule per amine and amidic proton*, and other solvation effects are captured using the SMD implicit solvation model. The amine and amide groups with intramolecular hydrogen bonding interactions are excluded from the addition of explicit solvent molecules, since the turns and twists formed by this interactions cannot accommodate an explicit water molecule. This avoids adding a random number of explicit water molecules that requires a proper equilibration and careful sampling of solvent molecules, potentially leading to substantial increases in the computational cost. In contrast, our approach is systematic and balanced, while keeping the computational costs low. These explicitly added water molecules were further geometry optimized at the B3LYP/6-31+G level of theory to obtain their best possible orientations while keeping the non-hydrogen atoms of the protein fixed to preserve the conformation. Here B3LYP/6-31+G method is used as a reasonably inexpensive method to get a good optimized structure incorporating hydrogen bonds with the explicitly added water molecules. This is expected to be more reliable than geometries obtained with MM or semiempirical methods.

3. Results and Discussion

3.1 Calibration of MIM2-NMR method *vs.* full calculation

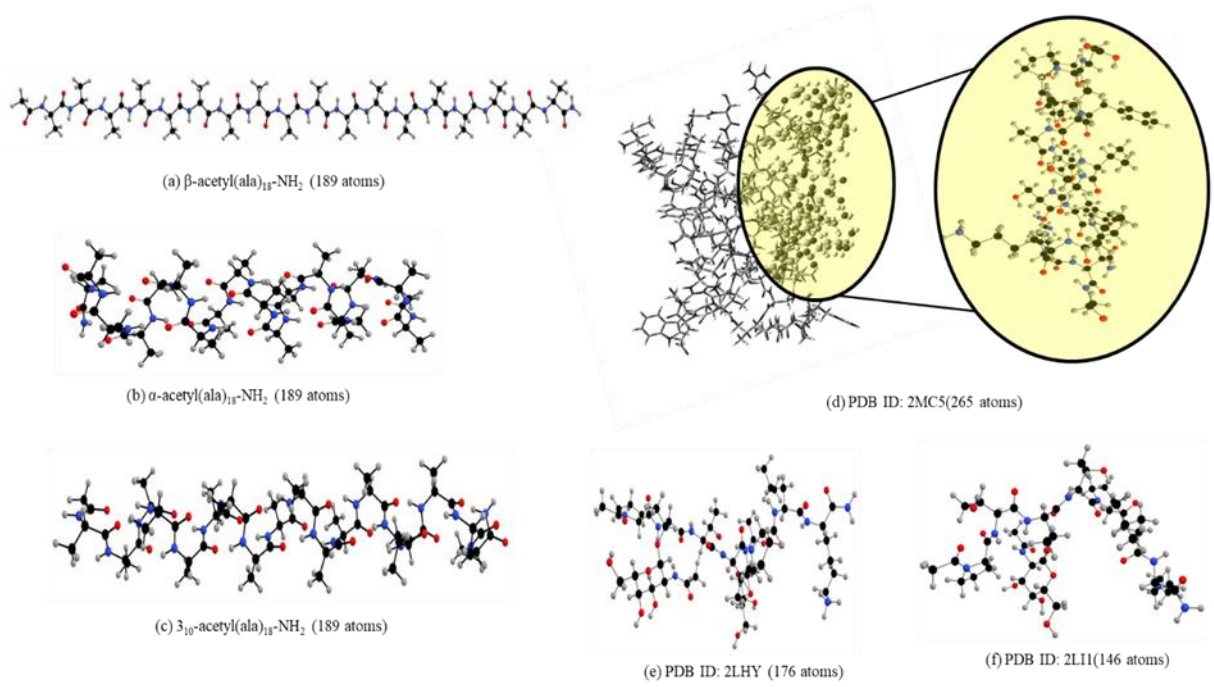

Figure 1. Molecules used to compare MIM method with the full molecule calculation. (a) β-acetyl-$(ala)_{18}$-NH_2 (189 atoms), structure label β-(Ala)18 , (b) α-acetyl-$(ala)_{18}$-NH_2 (189 atoms), structure label: α-(Ala)18, (c) 3_{10}-acetyl$(ala)_{18}$-NH_2 (189 atoms), structure label 3_{10}-(Ala)18, (d) α helical section highlighted in black circle is studied. PDB ID: 2MC5(265 atoms), structure label: 2MC5, (e) PDB ID: 2LHY (176 atoms), structure label: 2LHY, and (f) PDB ID: 2LI1(146 atoms), structure label: 2LI1.

The first part of our calibration is to assess the *impact of fragmentation*, i.e., how well does MIM-NMR perform in calculating chemical shifts with respect to the *full unfragmented calculations*. Only then can the performance of MIM-NMR with respect to experiment be assessed. A test set of six biomolecules, shown in **Figure 1**, was assembled to calibrate the performance of our MIM-NMR method. This set includes ***β***, ***α***, and ***3_{10}*** conformers of $(alanine)_{18}$ taken from our previous study,[100] and three proteins from protein data bank (PDB) comprising the *Escherichia coli* transcription protein (PDB ID **2MC5**; BMRB ID 19428)[101], and two sugar-binding, mucin glycoproteins (PDB IDs **2LHY** and **2LI1;** BMRB IDs 17871, and 17874, respectively). [102] The test set of biomolecules in this calibration have various intramolecular interactions such as backbone-backbone, backbone-side chain, and side chain-side chain hydrogen bonding networks that are commonly present in the majority of proteins. The *α* conformer of $(alanine)_{18}$ is selected

to account for tight helical turns with strong H-bonding interactions, whereas ***β*** and ***3₁₀*** conformers represent systems with comparatively weaker hydrogen bonding interactions. The proteins **2MC5**, **2LHY**, and **2LI1,** have both 3_{10} and 3_{14} helical turns, ***β*** strands, ***β*** bridges, and other connecting primary amino acid residues representing different intramolecular interactions present in diverse proteins. Additionally, the molecules in this test set also include mutated amino acid residues to test the performance of our protocol in the case of nonstandard residues. Overall, the variety of intramolecular interaction networks demands an appropriate fragment length to capture the primary interactions in the MIM2-NMR method. Based on initial exploratory studies, tetramer primary subsystems have been carefully explored in this study. As an example, for molecule 2LHY containing 7 backbone amino acid units along with two from mutated side chains (total of 9 amide groups), there are four primary tetramer subsystems and three derivative subsystems. In general, the number of the MIM subsystems will depend on the amino acid subunits present in the parent molecule and will grow *only linearly* with the size of the system.

To calibrate the performance of MIM2-NMR protocol, we first computed the ^{1}H, ^{13}C, and ^{15}N absolute chemical shifts of the test set proteins in the gas phase and compared with the conventional full system calculation performed at *m*PW1PW91/6-311G(d,p) level of theory, as shown in **Table 1.**

Table 1. Maximum deviation and mean absolute deviations for calculated chemical shifts at MIM2 [MPW1PW91/6-311G(d,p):MPW1PW91/6-31G(d)] compared to NMR chemical shift calculated for the full, unfragmented molecules in test set I.

		Maximum deviation			**Mean absolute deviation**			
system	NBasis	^{1}H	^{13}C	^{15}N	^{1}H	^{13}C	^{15}N	total
β (alanine)$_{18}$	2262	0.01	0.03	0.02	0.00	0.02	0.01	0.02
α (alanine)$_{18}$	2262	0.03	0.29	0.34	0.01	0.13	0.13	0.30
3_{10} (alanine)$_{18}$	2262	0.03	0.15	0.17	0.01	0.07	0.07	0.36
2MC5	3162	0.09	0.62	0.33	0.02	0.13	0.15	0.21
2LHY	2046	0.02	0.02	0.02	0.00	0.01	0.01	0.01
2LI1	1716	0.01	0.02	0.04	0.00	0.01	0.02	0.01
Average		0.03	0.19	0.15	0.01	0.06	0.08	0.15

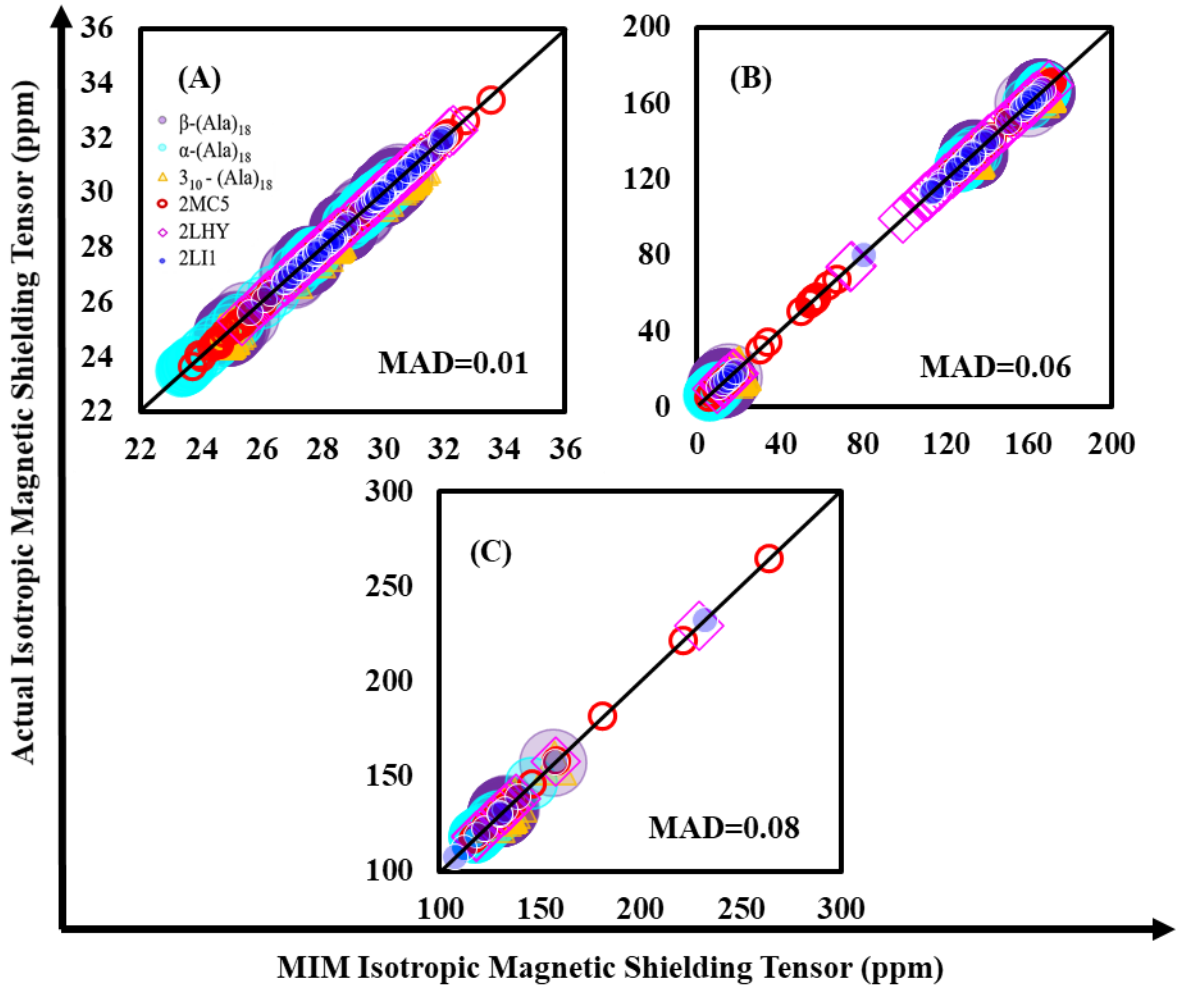

Figure 2. Comparison of (A) ^{1}H, (B) ^{13}C and (C) ^{15}N isotropic magnetic shielding tensors (ppm) in the $(Ala)_{18}$ conformers and protein structures from PDB, evaluated using full, unfragmented molecule at MPW1PW91/6-311G(d,p) level and MIM2 at [MPW1PW91/6-311G(d,p):MPW1PW91/6-31G(d)] level of theory. In each graph, structures are color coded as, β-$(Ala)_{18}$ in purple, α-$(Ala)_{18}$ in teal, 3_{10}-$(Ala)_{18}$ in yellow, 2MC5 in red, 2LHY in magenta and 2LI1 in blue.

The geometries for ***α***, ***β***, and ***3₁₀*** conformers of $(alanine)_{18}$ (henceforth **$(ala)_{18}$**) were obtained from the paper by Saha and Raghavachari,[100] whereas the geometries of **2LHY**, **2MC5** and **2LI1** were obtained from the PDB database without any further change. The NMR chemical shifts were calculated using MIM2[*m*PW1PW91/6-311G(d,p):*m*PW1PW91/6-31G(d)] and compared with the full system calculation (without any fragmentation), as shown in **Figure 2**.

The mean absolute deviations (MADs) in isotropic magnetic shielding tensors (in ppm) calculated using the MIM2 method were compared with the corresponding full calculations and are listed in **Table 1**. The average MAD for the proton (^{1}H), carbon (^{13}C), and nitrogen (^{15}N) chemical shifts are only 0.01, 0.06, and 0.08 ppm, respectively. These errors are only about 3% of our target accuracy in the calculated NMR chemical shifts (viz., errors of < 0.3 ppm for ^{1}H, < 2-3 ppm for ^{13}C, and < 3-4 ppm for ^{15}N) with a remarkably good correlation (R: 0.99-1.00) for all the NMR active nuclei. This comparison demonstrates that the MIM2-NMR method accurately reproduces the NMR spectra calculated for the full, unfragmented, protein molecule. It also allows us to extend our method for large systems without having to perform expensive, sometimes unaffordable, QM calculations on the full molecule. The remaining sections of the paper will be devoted to assessing the performance of MIM-NMR for calculating ^{1}H and ^{13}C chemical shifts with respect to experiment.

3.2 Development of a MIM2-NMR protocol for the prediction of NMR chemical shifts

To develop a reliable protocol for performing accurate NMR calculations using MIM, we chose two glycoproteins, **2LI1** and **2LHY**, that are essential for the antibody recognition in anticancer-vaccine developments.[102] These two molecules were selected mainly for the following reasons. Both **2LI1** and **2LHY** are relatively smaller proteins with only 8 and 9 amino acid residues, respectively, and have experimentally determined NMR spectra. The relatively small size of these proteins makes it possible to include the effects of multiple conformations, which may be necessary to identify and assign the NMR spectra (*vide infra*) correctly. Additionally, since the NMR-derived structures of **2LI1** and **2LHY** already include a total of 27 and 28 conformers, respectively, no further conformational search had to be performed for these molecules. As in many proteins, both **2LI1** and **2LHY** proteins have polar functional groups and side chains which give an overall charge to these proteins. Since the electrostatic interactions are overestimated substantially in the gas phase, to obtain a more reasonable stabilization appropriate for such species in solution, charged residues are neutralized. This approach has previously been shown to be a reasonable approximation.[40, 96]

3.2.1 Multiple Conformations and Boltzmann averaging

To check the performance of MIM2-NMR method with respect to experiments, the relative energies of the various NMR-derived conformers of both the **2LHY** and **2LI1** protein have been determined using four DFT methods (i) B3LYP-D3BJ/6-31+G(d) (ii) CAM-B3LYP-D3BJ/6-

31+G(d) (iii) *m*PW1PW91/6-31+G(d) (iv) *ω*B97X-D/6-31+G(d). To assess the conformational effects on NMR chemical shift predictions, no further post-processing was performed on the proteins. All four DFT methods consistently gave the conformer 9 as the lowest energy conformer for **2LHY** protein, and the energy of the second-lowest energy conformer is calculated to be 3 to 7 kcal/mol higher than conformer 9 (full results are given in **Tables S1-S4** of the supporting information). A pictorial representation of relative energies of the **2LHY** conformers is shown in **Figure 3**.

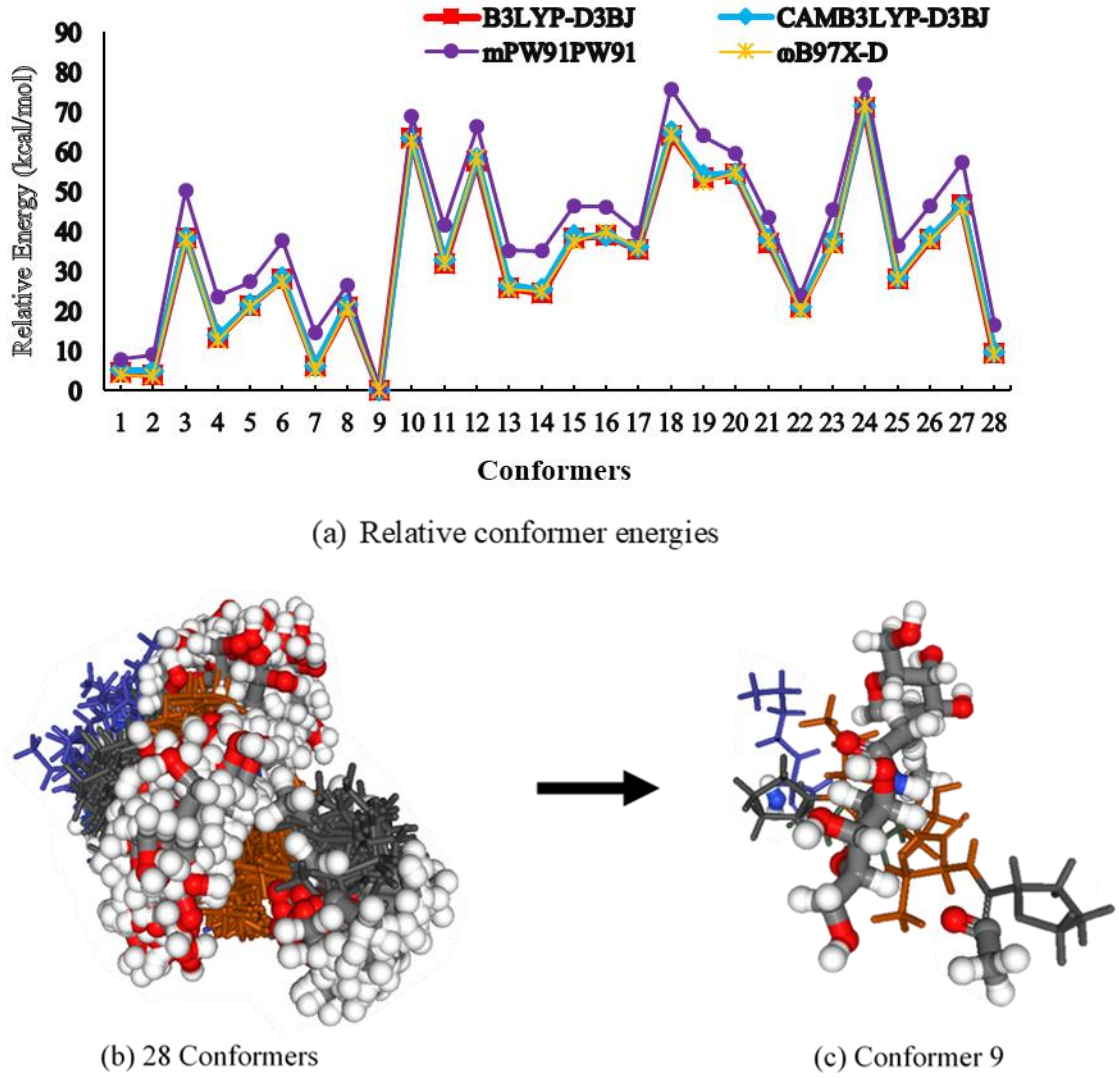

Figure 3 . Conformational analysis of various conformers of **2LHY** protein (a) relative energies calculated using B3LYP-D3BJ (red), CAM-B3LYP-D3BJ (blue), *m*PW91PW91 (violet), and *ω*B97X-D(yellow) DFT functionals and 6-31+G(d) basis set. (b) superposition of 28 conformers, and (c) structure of conformer 9.

On the other hand, in the case of **2LI1**, two conformers, 19 and 23, show significant (>10%) Boltzmann population contributions, while a third conformer, 1, shows a small contribution (~5%). However, the energy difference between conformers 19 and 23 of 2LI1 is calculated to be very small with all the methods (1 kcal/mol or less), and the lowest energy conformer is found to be sensitive to the DFT method used. Two of the four considered DFT methods, namely *m*PW1PW91 and ωB97X-D, gave conformer 23 as the lowest energy conformer, whereas the other two DFT methods (i.e., B3LYP-D3BJ and CAM-B3LYP-D3BJ) gave conformer 19 as the lowest energy conformer.

It is important to note that, in general, the experimentally observed NMR spectra may have contributions from a mixture of the low-lying conformations present in the sample. Nevertheless, many theoretical NMR spectral prediction methods commonly use only a single input conformer.[71, 103, 104] In principle, the lowest energy structures might not be enough to obtain an accurate spectrum. To assess this quantitatively, first we computed the NMR spectra using our MIM2-NMR protocol using only the lowest energy conformers of **2LHY** and **2LI1** proteins. Then, we computed the NMR spectra by including the contributions from other conformers as a Boltzmann average using Equation (6). For the MIM2-NMR calculations, four different combinations of DFT methods were considered: (i) *m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G, (ii) CAM-B3LYP-D3BJ/6-311++G(2d,2p):CAM-B3LYP-D3BJ/6-31G, (iii) B3LYP-D3BJ/6-311++G(2d,2p):B3LYP-D3BJ/6-31G, and (iv) ωB97X-D/6-311++G(2d,2p):ωB97X-D/6-31G.

Table 2. Mean Absolute Deviation (MAD) values of ^{1}H and ^{13}C NMR chemical shifts using MIM2[X/6-311++G(2d,2p) :X/6-31G] of **2LHY** protein with respect to experimental NMR chemical shifts. (X is different density functional methods for MIM2-NMR calculations.)

No.	MIM2-NMR THEORY		CONFORMER 9		
			^{1}H	^{13}C	Total
1	B3LYP-D3BJ	MAD	0.93	3.19	1.98
		R	0.86	0.99	
2	CAMB3LYP-D3BJ	MAD	0.89	2.82	1.80
		R	0.87	0.99	
3	ωB97XD	MAD	1.24	3.03	2.10
		R	0.78	0.99	
4	*m*PW1PW91	MAD	0.84	2.67	1.78
		R	0.88	0.99	

Table 3 . Mean Absolute Deviation (MAD) values of ^{1}H and ^{13}C NMR chemical shifts using MIM2[X/6-311++G(2d,2p) :X/6-31G] of **2LI1** protein with respect to experimental NMR chemical shifts. (X is different density functional methods for MIM2-NMR calculations.)

No.	MIM2-NMR THEORY		CONFORMER 19			CONFORMER 23		
			^{1}H	^{13}C	Total	^{1}H	^{13}C	Total
1	B3LYP-D3BJ	MAD	0.80	3.07	1.79	0.89	2.72	2.44
		R	0.90	0.99		0.88	0.99	
2	CAMB3LYP-D3BJ	MAD	0.81	3.17	1.82	0.90	2.51	3.35
		R	0.90	0.99		0.88	0.99	
3	ωB97XD	MAD	0.78	3.03	1.76	0.87	2.42	3.12
		R	0.90	0.99		0.88	0.99	
4	*m*PW1PW91	MAD	0.78	2.97	1.74	0.89	2.47	2.93
		R	0.90	0.99		0.88	0.99	

The MIM2-NMR results calculated for the lowest energy conformers of **2LHY** (conformer 9) and **2LI1** (conformers 19 and 23) are shown in **Tables 2** and **3**, respectively. We note that solvation effects (*vide infra*) are not included in these initial results.

Our calculation shows that conformer 9 of the **2LHY** protein with the *m*PW1PW91 method showed the lowest MAD value of 0.84 ppm for ^{1}H and 2.67 ppm for ^{13}C chemical shifts. (**Table 2**). Similar results are obtained for the conformer 19 of the **2LI1** protein with the *m*PW1PW91 method (MAD values of 0.78 ppm for ^{1}H and 2.97 ppm for ^{13}C) (**Table 3**). For the ωB97X-D method, although the MAD in the calculated chemical shifts for ^{1}H and ^{13}C is comparable to the results for *m*PW1PW91 for **2LI1**, somewhat larger deviations are observed in the case of **2LHY**. Interestingly, the conformer 23 of **2LI1** protein that was calculated to be the lowest energy conformer by two DFT methods, showed the larger MAD values for ^{1}H NMR, as shown in **Table 3**.

The effect of the different conformations on the NMR spectra can be explored for **2LI1.** As mentioned earlier, **2LI1** protein has three conformers (conformers 19, 23, and 1) within the 3 kcal/mol energy window and could play a significant role in obtaining accurate NMR spectra. The relative abundance of all conformers of **2LI1** protein is shown in **Figure 4**. To explore the effect of including multiple conformations in the NMR calculations, we computed the NMR chemical

shifts using the abovementioned four MIM2-NMR methods. Due to the small difference in the relative energies among the conformers, the different methods gave different weighted contributions of the conformer abundance (shown in **Tables S5-S8** of the supporting information).

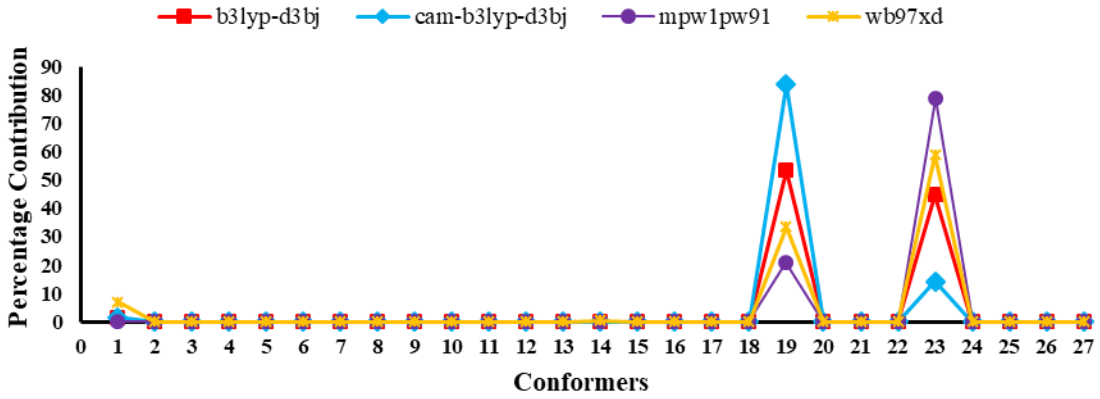

Figure 4 . Relative energies of 27 conformers of 2LI1 protein from protein data bank (PDB) expressed as Boltzmann populations. Red: B3LYP-D3BJ. blue: CAMB3LYP-D3BJ. violet: mPW91PW91. Yellow: ωB97X-D. All energies determined as single points using the 6-31+G* basis set.

We have calculated the Boltzmann averaged ^{1}H, and ^{13}C MAD values for different combinations of the DFT methods used in MIM2-NMR calculations with respect to the experimental values (full results are given in **Table S9** of the supporting information). Overall, considering the entire range of chemical shifts for NMR active nuclei in **2LI1** protein, the NMR chemical shifts evaluated using the combination of MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] method for the NMR chemical shift calculation that are weighted using the conformer populations obtained using the CAM-B3LYP/6-31+G(d) method results in the smallest MAD value with the best correlation with respect to the experiments. This protocol gave the best MAD values of 0.76 ppm for ^{1}H and 2.74ppm for ^{13}C. These values are very slightly improved from the corresponding values obtained using the most stable isomer 19 (MAD values of 0.78ppm for ^{1}H and 2.97ppm for ^{13}C).

3.2.2 Solvation effects and optimization of the MIM2-NMR protocol

It is well known that accurate prediction of ^{1}H NMR is challenging compared to other NMR active nuclei in the proteins. Hydrogen bonding interactions play an important role in determining the structure of a protein and are mostly influenced by the hydrogens in the system. Thus, we use the improvement in proton chemical shifts to optimize our protocol for the analysis of chemical shifts of both **2LHY** and **2LI1** proteins. From our analysis in the previous section,

combination of CAM-B3LYP-D3BJ/6-31+G(d) to analyze abundance and MIM2[*m*PW1PW91/6-311++ G(2d,2p):*m*PW1PW91/6-31G] method for NMR chemical shifts, gives the least MAD value in chemical shift for ^{1}H and ^{13}C in both **2LHY** and **2LI1** proteins.

We have evaluated the MIM2-NMR spectra for **2LHY** and **2LI1** proteins using 4 different models representing different protocols for the inclusion of solvent effects, and the results are shown in **Figures 5-8**. Briefly, the 4 models correspond to (**A**) MIM2-NMR calculated in the gas-phase (MIM_{gas}) without any structure minimization, (**B**) MIM2-NMR calculated in the gas-phase using MM minimized structure ($MIM_{gas}^{restraint}$), (**C**) MIM2-NMR calculated with implicit solvation only ($MIM_{implicit}$), and (**D**) MIM2-NMR with the explicit-implicit solvation model ($MIM_{explicit-implicit}^{restraint}$). More details of the 4 models and their performance are discussed below.

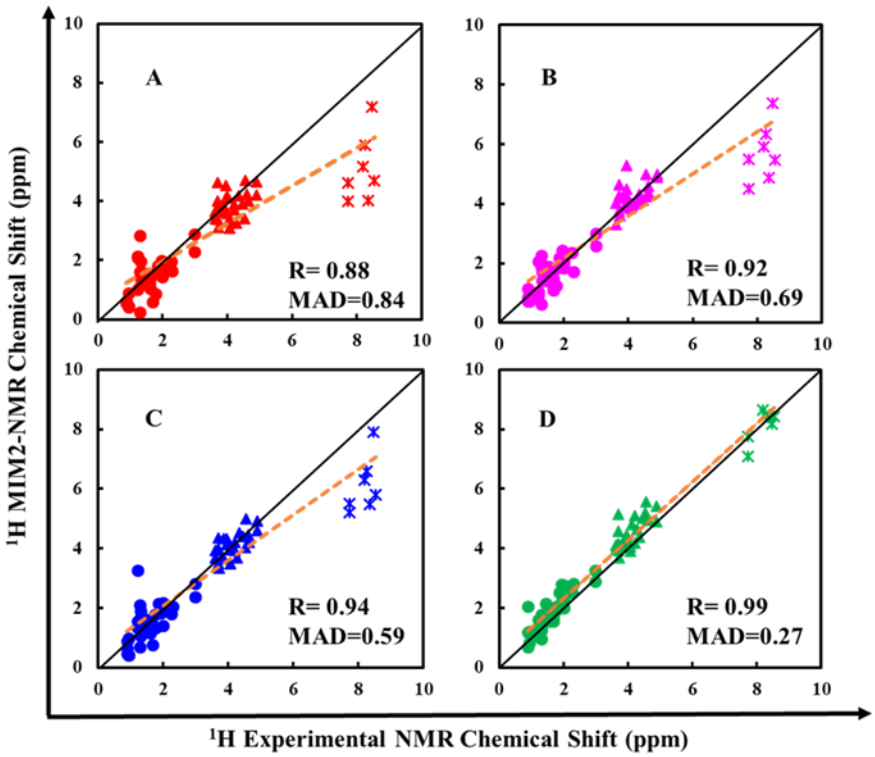

Figure 5. Comparison of experimental ^{1}H NMR of sugar binding protein **2LHY** with MIM2 ^{1}H NMR calculated at MIM2[*m*PW1PW91/6-311++G(2d,2p): *mPW*1PW91/6-31G] level in (A) MIM_{gas}, (B) $MIM_{gas}^{restraint}$, (C) $MIM_{implicit}$ and (D) $MIM_{explicit-implicit}^{restraint}$. The MIM-calculated ^{1}H NMR chemical shifts are depicted with reference to Tetramethyl silane (TMS). (Asterisk markers for amidic protons, triangle markers for α ^{1}H 's and circle markers for the rest of the protons are used in the plot.)

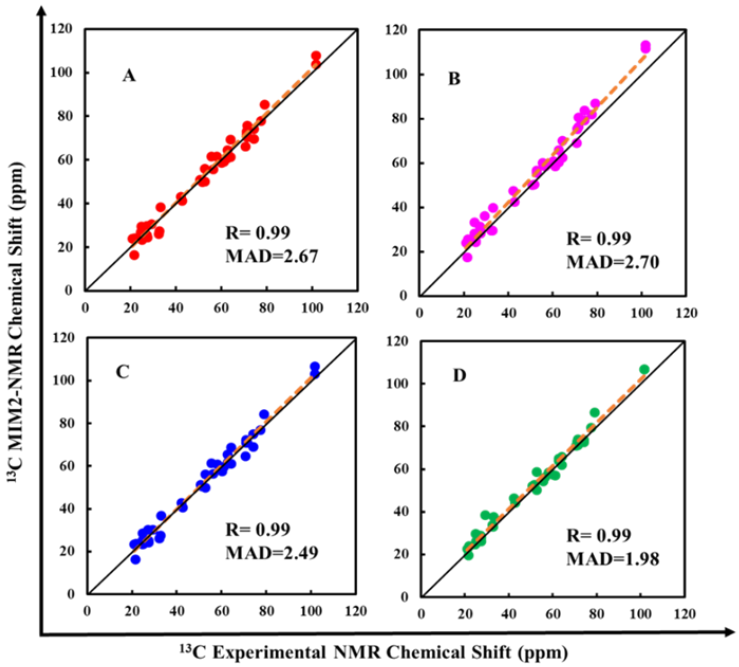

Figure 6. Comparison of experimental ^{13}C NMR spectrum of Sugar binding protein **2LHY** molecule with neutral residues using (A) MIM_{gas}, (B) $MIM_{gas}^{restraint}$, (C) $MIM_{implicit}$ and (D) $MIM_{explicit-implicit}^{restraint}$ models at the MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] level. The MIM ^{13}C NMR chemical shifts are depicted in reference to Tetramethyl silane (TMS).

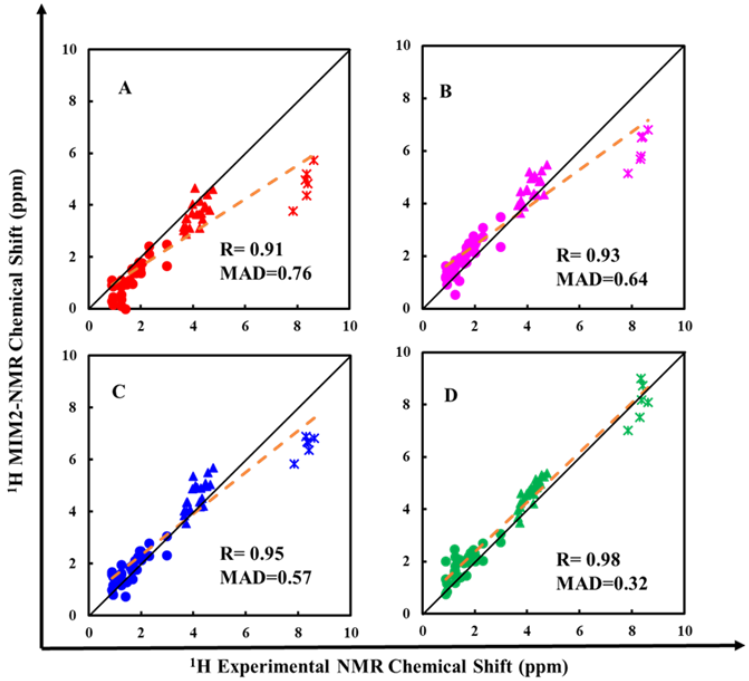

Figure 7. Comparison of experimental ^{1}H NMR of sugar binding protein **2LI1** with MIM2 ^{1}H NMR calculated at MIM2[*m*PW1PW91/6-311++G(2d,2p): *mPW*1PW91/6-31G] level in (A) MIM_{gas}, (B) $MIM_{gas}^{restraint}$, (C) $MIM_{implicit}$ and (D) $MIM_{explicit-implicit}^{restraint}$. The MIM-calculated ^{1}H NMR chemical shifts are depicted with reference to Tetramethyl silane (TMS). (Asterisk markers for amidic protons, triangle markers for α ^{1}H's and circle markers for the rest of the protons are used in the plot.)

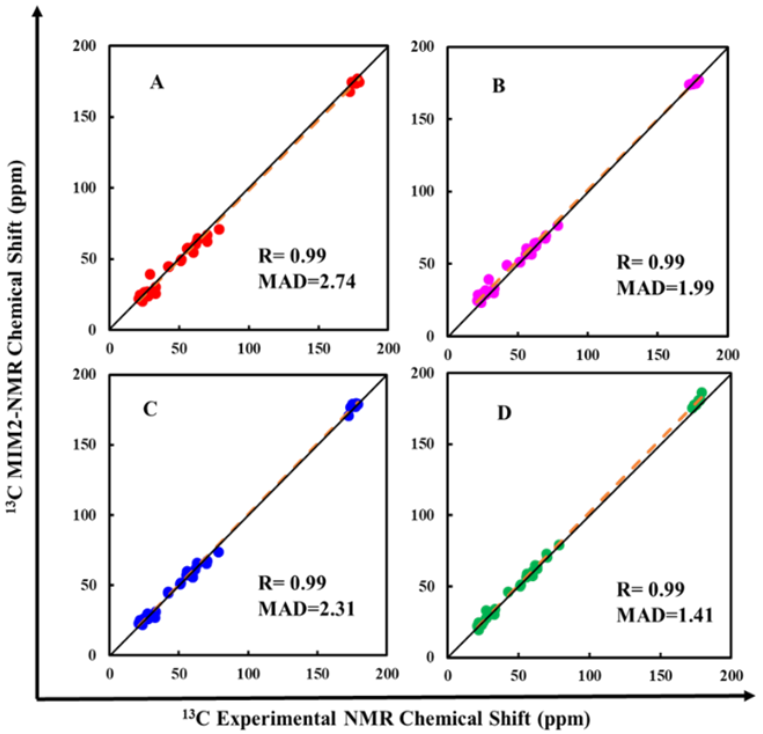

Figure 8. Comparison of experimental ^{13}C NMR spectrum of Sugar binding protein **2LI1** molecule with neutral residues using (A) $MIM2_{gas}$, (B) $MIM2^{restraint}_{gas}$, (C) $MIM2_{implicit}$ and (D) $MIM2^{restraint}_{explicit-implicit}$ models at the MIM2[mPW1PW91/6-311++G(2d,2p):mPW1PW91/6-31G] level. The MIM ^{13}C NMR chemical shifts are depicted in reference to Tetramethyl silane (TMS).

3.2.2.1 Gas phase MIM2-NMR calculations

Figures 5A and **7A** show the comparison of experiment with MIM2-Boltzmann-weighted-*gas-phase* (MIM_{gas}) NMR chemical shifts of ^{1}H's for **2LHY** and **2LI1** proteins, respectively. The linearly fitted plots for both proteins show large deviations in the range of 6-10 ppm where amidic protons are seen (indicated by the circle markers). Note that the α ^{1}H's (triangle markers in **Figure 5A** and **Figure 7A**) and the rest of the protons in the system (asterisk markers in **Figure 5A** and **Figure 7A**) show a good agreement with the experimental chemical shifts. MIM_{gas} NMR shifts of ^{13}C, as displayed in **Figures 6A** and **8A**, show an excellent agreement with experimental values with a correlation coefficient of 0.99 for both **2LHY** and **2LI1** proteins.

3.2.2.2 Molecular Mechanics (MM) restraint minimized MIM2-NMR calculations

To assess the effect of geometry optimization on the accuracy of calculated NMR spectra, the lowest energy structures were minimized using Molecular Operating Environment (MOE) with the AMBER10:EHT force field.[105, 106] A range of restraint parameter values ranging from 0.5 Å

to 2.0 Å was set for every atom in the proteins, and the effect on predicting the chemical shifts using MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] method was analyzed (full results are given in **Tables S10** and **S11** of the supporting information). A restraint optimization parameter value of 0.5 Å resulted in the lowest MAD value of 1.48 ppm for all NMR active nuclei of **2LHY** protein compared to other parameters set for MM restraint minimization. For **2LI1** protein, comparable NMR results were obtained for the 0.5 Å and 1.0 Å constraint optimized structures. Since the 0.5 Å restraint-optimized structure resulted in overall good results, we employed this parameter to obtain the optimized geometry for the following analysis. As shown in **Figures 5-8B**, the MIM gas phase restraint optimization structures ($\mathrm{MIM}_{\mathrm{gas}}^{\mathrm{restraint}}$) resulted in a significant improvement in calculated chemical shifts compared to the results for the unoptimized structures.

3.2.2.3 Implicit and explicit-implicit solvation model for MIM2-NMR calculations

Although we have seen a significant improvement in the accuracy in the calculated NMR chemical shifts for ^{1}H and ^{13}C, the results are still far from our target accuracy (0.3 ppm for ^{1}H and 2-3 ppm for ^{13}C). In particular, larger deviations are seen for NMR active nuclei like amine and amide ^{1}H's, which are more susceptible to the solvent environment due to the possibility of forming hydrogen bonding interactions with solvent water molecules. This suggests that a proper accounting of the solvation effect and further energy minimization may be necessary to lower the errors. Adding the first solvation shell water molecules (waters within 3 Å of the protein) to **2LHY** gave a total of 488 atoms with 104 water molecules, and for **2LI1** protein, a total of 395 atoms with 83 water molecules. However, constrained optimization of the solvated protein in external water molecules will be needed to obtain a reasonable starting structure to compute the chemical shifts. Contrary to implicit solvation which negligibly affects the computational cost, energy minimization of explicitly solvated protein is computationally intensive and can be a limiting step for calculating the NMR shielding tensors. In order to account for the solvation effect, we have modeled the aqueous solvent environment through the implicit solvation (SMD solvation model), and a modified combination of implicit and explicit solvation models (micro-solvation approach) with less computational cost as described in the Methods section. In particular, *a single water molecule per amine and amide units* were found to be sufficient to improve the performance substantially. As mentioned earlier, amine and amide groups with intramolecular hydrogen

bonding interactions are left as such without any explicit water molecule in our micro-solvation approach.

For **2LHY** and **2LI1** MM restraint minimized structure, a total of 10 and 6 explicit water molecules were added near the ^{1}H atoms attached to nitrogen (both amine and amide protons) and a MIM2-NMR calculation with the explicit-implicit solvation ($MIM^{restraint}_{explicit\text{-}implicit}$) calculation is performed in SMD implicit solvation to predict the NMR chemical shifts. Improvement of ^{1}H NMR chemical shifts in the entire range of chemical shifts for **2LHY** protein is depicted in **Figure 5A-D.** (**5A**) shows the results of MIM2-NMR calculated in the gas-phase (MIM_{gas}) without any structure minimization, (**5B**) shows the results of MIM2-NMR calculated in the gas-phase using MM restraint minimized structure ($MIM^{restraint}_{gas}$), (**5C**) shows the results obtained for MIM2-NMR calculated with implicit solvation only ($MIM_{implicit}$), and (**5D**) shows the results for MIM2-NMR with the explicit-implicit solvation model ($MIM^{restraint}_{explicit\text{-}implicit}$). Comparing the results obtained for the four computational protocols used, a systematic improvement can be observed. The MAD values in ^{1}H NMR chemical shifts improved from 0.84 ppm (**Figure 5A**) to 0.69 ppm (**Figure 5B**) when MM-optimized structure was used to calculate the MIM2-NMR in the gas phase instead of raw structure obtained from PDB. The effect of implicit solvation was seen to reduce the calculated MAD value of ^{1}H NMR by 0.10 ppm units (MAD = 0.59 ppm, **Figure 5C**). More dramatic improvement was observed with the explicit-implicit solvation model, which lowered the MAD value by more than half, yielding 0.27 ppm deviation from experiment. Along with the improvement seen in ^{1}H MAD chemical shift values, the correlation coefficient (R) also improved quite remarkably from a value of 0.88 in the gas phase to 0.99 with the explicit-implicit solvation.

For ^{13}C chemical shifts, the results from various MIM2-NMR models and their comparison with experimental chemical shifts for **2LHY** are depicted in **Figure 6**. For ^{13}C chemical shifts, although a good correlation (R = 0.99) is observed even in the absence of any solvation (which remains the same with the solvation effect included), the overall MAD value improves from 2.67 ppm calculated in the gas phase to 1.98 ppm with the final explicit-implicit solvation model. To visualize the errors more fully, individual plots of error vs. shielding for ^{13}C chemical shifts are shown in **Figure S1** of the supporting information.

The trends observed using the different computational models of MIM2-NMR calculations for **2LI1** protein are quite similar to the results obtained for **2LHY** protein. For ^{1}H, MIM-NMR

models are compared with experimental chemical shifts, as shown in **Figure 7A-D**. MAD values for entire range of ^{1}H are (**7A**) 0.76, (**7B**) 0.64, (**7C**) 0.57 and (**7D**) 0.32 ppm with correlation coefficients of (**7A**) 0.91, (**7B**) 0.93, (**7C**) 0.95 and (**7D**) 0.98 respectively. **Figure 8A-D** depicts the ^{13}C correlation graphs of MIM2-NMR versus experiments, and shows a MAD of (**8A**) 2.74, (**8B**) 1.99, (**8C**) 2.31, and (**8D**) 1.41 ppm with a correlation coefficient of 0.99 for all computational models (error vs shielding plot is shown in **Figure S2** of the supporting information).

Since the Boltzmann-averaged structure of **2LI1** comprises of three different conformers, we have made a final comparison between the MAD values of lowest energy conformer and the Boltzmann averaged results of **2LI1** protein calculated using the $MIM^{restraint}_{explicit\text{-}implicit}$ method. For the lowest energy conformer (conformer 19), the calculated MAD value are 0.36 and 1.72 ppm for ^{1}H and ^{13}C respectively. The MAD values of the lowest energy conformer are improved for the Boltzmann averaged structure yielding the MAD values of 0.32 ppm and 1.41 ppm for ^{1}H and ^{13}C **(Figures 7D** and **8D**). It is interesting to note that the dominant isomer for **2LI1** with a 84% Boltzmann weighted chemical shift comes from conformer 19.

3.3 Application of the MIM2-NMR protocol for the prediction of NMR chemical shifts

The performance of the MIM2-NMR protocol calibrated above has been assessed on two other standard, but larger, proteins: PDB IDs **2MC5**[101] (BMRB 19428) and **3UMK**[107]. For **2MC5,** which has a single conformer submitted in PDB, a 17-residue slice beginning from residue number 46 to 62 (a total of 265 atoms) from the NMR-derived protein structure was used. A 33-residue slice with the residues starting from 535 to 567 (a total of 547 atoms) of solution NMR-derived **1TKN**[108] (BMRB 6236) is used for chemical shift assessments using the structural coordinates obtained from X-ray crystallographic structure **3UMK**. For **2MC5**, a total of 11 water molecules was added externally, forming hydrogen bonds with the protein along with the SMD implicit solvation to model the solvent environment. Similarly, for **3MUK**, a total of 15 explicit water molecules along with the implicit solvation model was employed. As noted earlier, in the case of both 2MC5 and 3UMK proteins, the explicit water molecules have been added near the exposed amine and amide groups with no intramolecular hydrogen bonding interactions.

Figures 9-10 show the linearly fitted correlation of MIM2-NMR computed chemical shifts of ^{1}H and ^{13}C with respect to the experimental values for the single conformer of the **2MC5** protein.

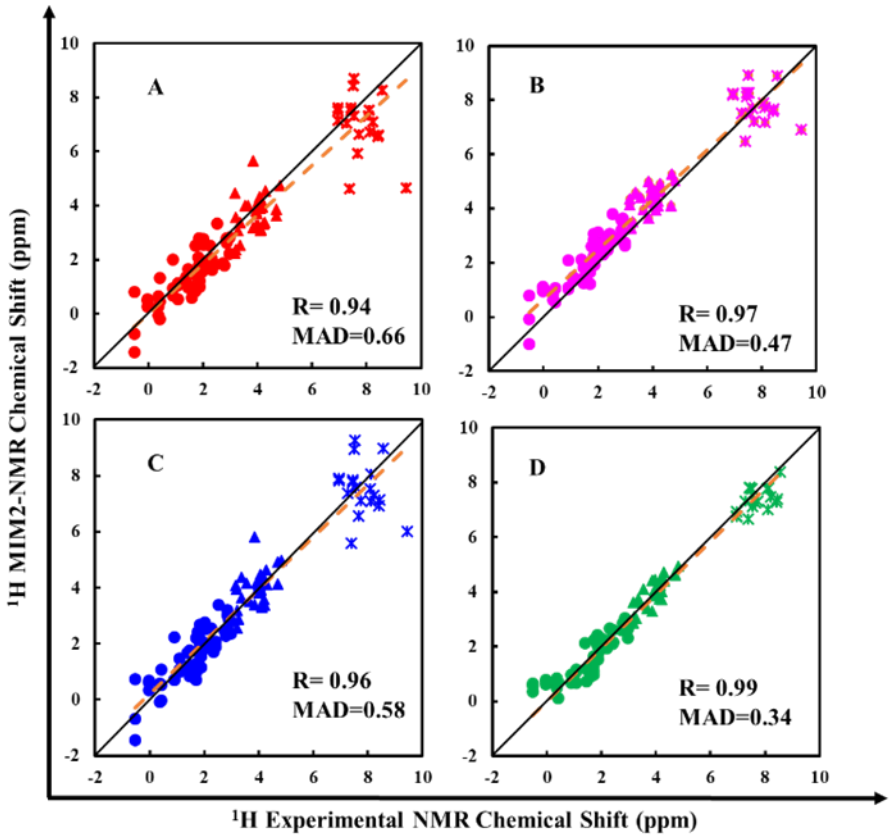

Figure 9. Comparison of experimental ^{1}H NMR of **2MC5** (residue numbers 46 to 62) with MIM2 ^{1}H NMR calculated at MIM2[mPW1PW91/6-311++G(2d,2p): mPW1PW91/6-31G] level in (A) MIM_{gas}, (B) $MIM_{gas}^{restraint}$, (C) $MIM_{implicit}$ and (D) $MIM_{explicit-implicit}^{restraint}$. The MIM-calculated ^{1}H NMR chemical shifts are depicted with reference to tetramethylsilane (TMS). (Asterisk markers for amidic protons, triangle markers for α ^{1}H's and circle markers for the rest of the protons are used in the plot).

As in the benchmarking study discussed above, panels **A-D** show the NMR results calculated using MIM_{gas}, $MIM_{gas}^{restraint}$, $MIM_{implicit}$, and $MIM_{explicit-implicit}^{restraint}$ models, respectively. For the ^{1}H NMR chemical shifts (**Figure 9**) calculated in the gas phase without further minimizing the PDB structure, we obtained a decent correlation (R = 0.94) with MAD value of 0.66 ppm. A significant improvement was observed when the MM-minimized structure was used to calculate ^{1}H NMR chemical shifts. The calculated MAD for the MM minimized structure is improved to 0.47 ppm (R = 0.97). This demonstrates that the restraint-minimization of the structure results in a smaller deviation in the calculated chemical shift values with a slightly better correlation with experiment. Surprisingly, the impact of including the implicit solvation effect on the computed NMR chemical shifts was found to lead to a small deterioration in the accuracy (MAD of 0.58 ppm, R = 0.96). However, substantial improvement (MAD of 0.34 ppm, R = 0.99) was seen when the solvation

effects were included using SMD explicit-implicit solvation on the MM minimized structure (MAD = 0.34 ppm; R = 0.99). This is very close to the target accuracy of 0.30 ppm. For ^{13}C NMR chemical shifts (**Figure 10**), the MAD values reduced from 2.35 ppm for the gas phase calculations to 2.12 ppm when the structure was minimized using MM. The solvation effects only yielded a small further improvement while maintaining a remarkable correlation (R = 0.99). For ^{13}C NMR chemical shifts, MAD values for the four computational models are (**A**) 2.35, (**B**) 2.12, (**C**) 1.89, and (**D**) 1.94 ppm with a correlation coefficient of 0.99 for each of the models (error vs shielding plot is shown in **Figure S3** of the supporting information). These values are well within our target accuracy of 2-3 ppm.

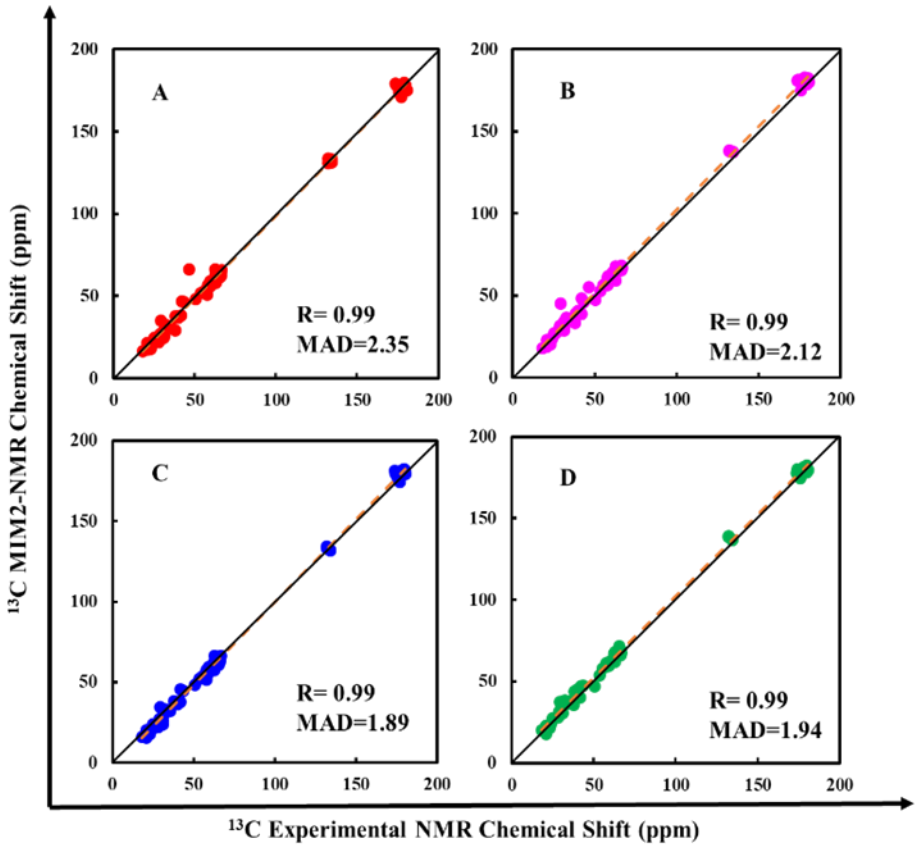

Figure 10. Comparison of experimental ^{13}C NMR of **2MC5** (residue numbers 46 to 62) with MIM2 ^{13}C NMR calculated at MIM2[*m*PW1PW91/6-311++G(2d,2p): *mPW*1PW91/6-31G] level in (A) MIM_{gas}, (B) $MIM_{gas}^{restraint}$, (C) $MIM_{implicit}$ and (D) $MIM_{explicit-implicit}^{restraint}$. The MIM-calculated ^{13}C NMR chemical shifts are depicted with reference to tetramethylsilane (TMS).

Similar improvements can be observed in the accuracy of MIM calculated NMR chemical shifts for the X-ray crystallography-derived molecule, **3UMK** (**Figures 11-12**).

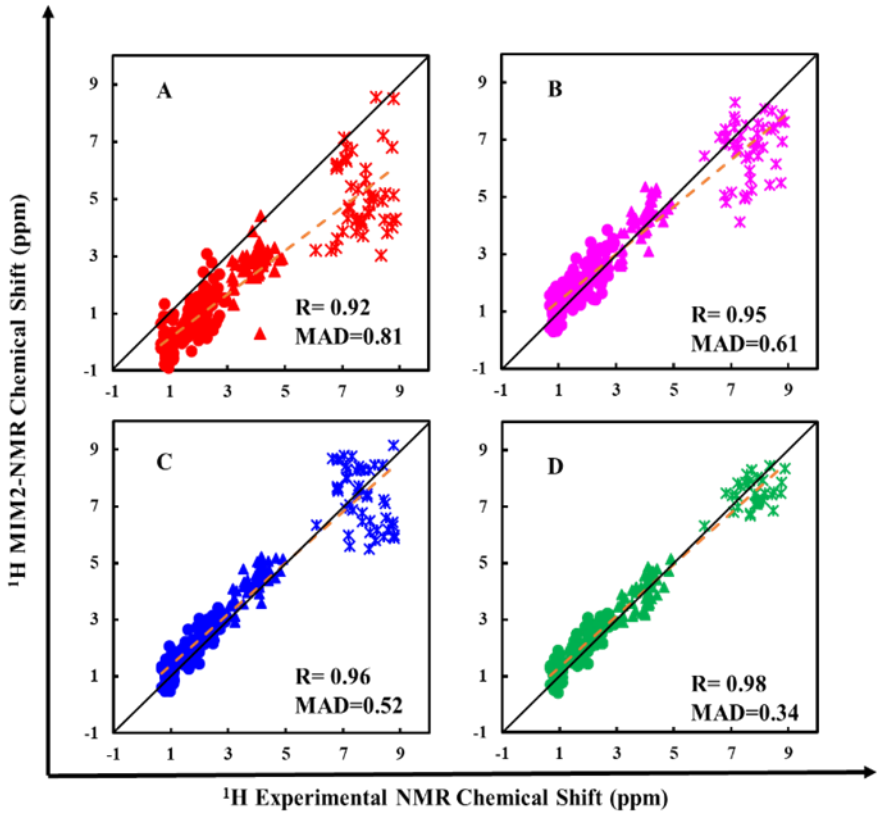

Figure 11. Comparison of experimental ^{1}H NMR of **3UMK** (residue number 535 to 567) with MIM2 ^{1}H NMR calculated at MIM2[*m*PW1PW91/6-311++G(2d,2p): *mPW*1PW91/6-31G] level in (A) MIM_{gas}, (B) $MIM_{gas}^{restraint}$, (C) $MIM_{implicit}$ and (D) $MIM_{explicit-implicit}^{restraint}$. The MIM-calculated ^{1}H NMR chemical shifts are depicted with reference to tetramethylsilane (TMS). (Asterisk markers for amidic protons, triangle markers for α ^{1}H's and circle markers for the rest of the protons are used in the plot.)

In this case, the error in ^{1}H NMR calculated using MIM lowered by more than a factor of 2, while going from unrefined gas-phase calculation (MAD = 0.81 ppm) to the MM geometry minimized structure in solution phase with explicit-implicit solvation (MAD = 0.34 ppm) along with a substantial improvement in the correlation coefficient (R improved from 0.92 to 0.98) (**Figure 11**). Similarly**, Figure 12** depicts the ^{13}C correlation of MIM2-NMR calculated under the abovementioned four solvation environments versus the experiment. As expected, the best results are obtained while using the MM-minimized structure with an explicit-implicit solvation model with a MAD value of 2.01 ppm with an excellent correlation of 0.99 (error vs shielding plot is shown in **Figure S4** of the supplementary information).

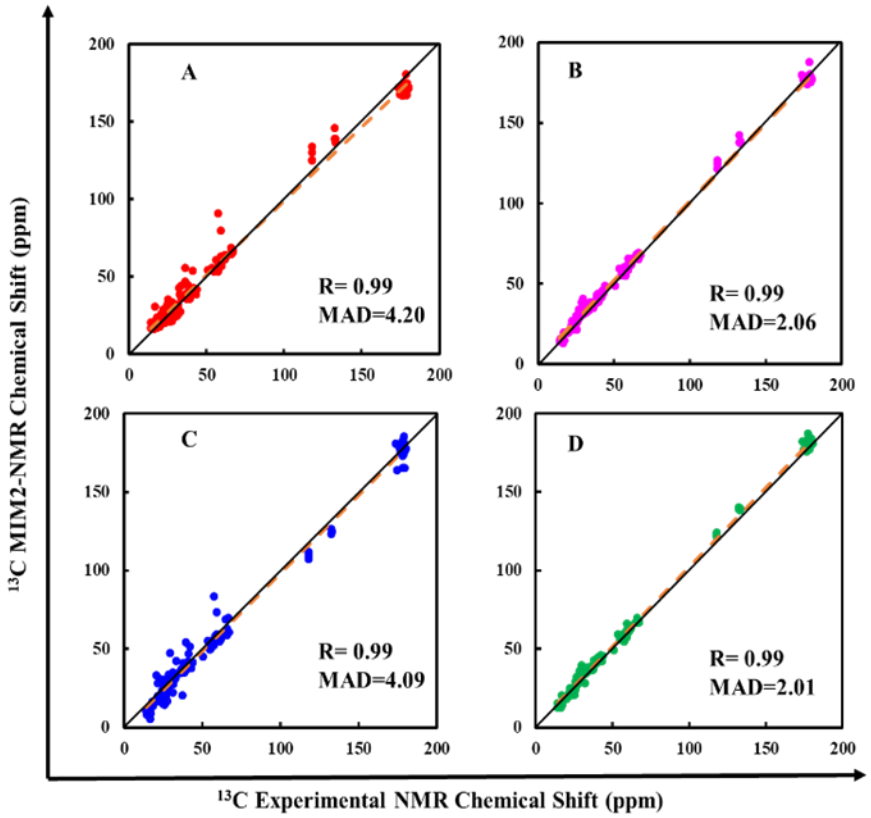

Figure 12 . Comparison of experimental ^{13}C NMR of **3UMK** (residue number 535 to 567) with MIM2 ^{13}C NMR calculated at MIM2[*m*PW1PW91/6-311++G(2d,2p): *mPW*1PW91/6-31G] level in (A) MIM_{gas}, (B) $MIM_{gas}^{restraint}$, (C) $MIM_{implicit}$ and (D) $MIM_{explicit-implicit}^{restraint}$. The MIM-calculated ^{13}C NMR chemical shifts are depicted with reference to tetramethylsilane (TMS).

Overall, our results show that the initial geometry optimization of database-harvested structures (NMR or X-ray derived) is necessary to lower the deviation from the experiment. Additionally, our calculations show that the implicit solvation alone is not sufficient to obtain the desired accuracy (~0.30 ppm for ^{1}H and ~2.0 ppm for ^{13}C). We also show that the use of just a few explicit solvent molecules to capture the local effects and implicit solvation to include the bulk effect in the calculated NMR provide reasonably accurate results. This is a highly effective way to include the solvation effects and improve the performance while keeping the computational costs low. Calculations using this new protocol are substantially faster than those with the inclusion of the complete first solvation shell of explicit water molecules.

For the peptides considered in this work, the most time-consuming components of MIM-NMR involve the primary subsystems. In general, the number of primary subsystems grows linearly with the size of the peptide while the size of the subsystem (tetrapeptide) is independent of the size of the parent molecule. Thus the scaling is linear and the speedup relative to the full (unfragmented) calculation increases with system size. Moreover, the calculations on all the subsystems can be done in parallel. In the case of MIM2 for very large molecules, the low level calculation on the unfragmented molecule can become rate limiting.

The advantage of MIM can be illustrated for the largest peptide that we have considered, the 33 peptide slice of 3UMK. The full molecule has 547 atoms, and the inclusion of 15 explicit water molecules increases the size to 592 atoms. Using the 6-311++G(2d,2p) basis set, this involves 10,790 basis functions, and a direct NMR calculation on the full molecule with this basis set is not feasible with our computational resources. Using MIM, however, the largest primary subsystem (tetrapeptide) involves only 90 atoms and 1,631 basis functions with the 6-311++G(2d,2p) basis set. This is the most expensive component of MIM1, but is easily accessible computationally. The corresponding low-level (6-31G basis set) calculations for MIM1 take negligible computer time. For MIM2, an additional calculation for the full molecule (547 atoms) with the 6-31G basis set involves 3,000 basis functions, and is also accessible. Thus MIM makes it possible to study these and larger peptide systems for accurate NMR chemical shift predictions. A Table containing computational timings for this system is included in the supporting information. (**Table S12**)

3.4 Comparison of MIM2-NMR results with SHIFTX2 and AF-QM/MM methods

Mean absolute deviation values of ^{1}H and ^{13}C chemical shifts for the **2LHY** protein obtained from the SHIFTX2 method can be compared with those from the MIM2-NMR method. SHIFTX2 predicted the NMR chemical shift of 39 out of 76 experimentally assigned protons of **2LHY** with a MAD of 0.08 ppm and R of 0.99, whereas MIM2-NMR gives a MAD value of 0.27 ppm with R of 0.99 for *all* 76 experimentally assigned protons. When the results for the same 39 protons was compared for both methods, MIM2-NMR gave a slightly worse mean absolute deviation of 0.22 (R = 0.99) compared to the SHIFTX2 results. Similarly, for ^{13}C, SHIFTX2 gave MAD of 1.51 ppm for 26 ^{13}C nuclei, and for the same nuclei MIM2-NMR gave a MAD of 1.99 ppm. However, MIM2-NMR predicts a MAD of 1.98 for *all* 40 ^{13}C NMR active nuclei of **2LHY**.

This comparison shows that in terms of accuracy, the results obtained using our MIM2-NMR method are slightly worse than the SHIFTX2 results. However, it is important to note that MIM2-NMR predicts the chemical shift values for *all the NMR active nuclei* while SHIFTX2 predicted only 65% of the reported experimental chemical shifts. This shows that the first-principles methods like MIM2-NMR may have significant advantages in predicting the chemical shifts of nonstandard chains in proteins, as seen in the case of the **2LHY** protein.

Swails *et al*. applied the AF-QM/MM method to calculate the NMR spectra of all residues of 2MC5 protein.[36] In their study, by excluding amide hydrogen atoms, they obtained an RMSE of 0.53 ppm for ^{1}H with correlation coefficient (R) of 0.97. They obtained the RMSE of 6.23 ppm for ^{13}C with R of 0.99, excluding all of the alpha carbon atoms.[36] In our study for a subset of residues of 2MC5, as discussed above (**Figures 9-10**), MIM2-NMR clearly shows a significant improvement for the error values of computed chemical shifts relative to experimental values.

4 Conclusions

In this work, we present an accurate MIM2-NMR method for the prediction of chemical shifts for large protein molecules. The MIM2-NMR method is calibrated using a collection of six polypeptides with the total number of basis functions ranging from 2262 to 3162 with 189 to 265 atoms. For comparison with the full unfragmented calculations, MIM2-NMR resulted in a MAD value of 0.01 ppm for ^{1}H, 0.06 ppm for ^{13}C, 0.08 ppm for ^{15}N, showing that the errors from fragmentation are very small (~3%) relative to of our target accuracy (*vide supra*). Evaluating the MIM2-NMR protocol with four different functionals for 2LHY protein showed that [*m*PW1PW91/6-311++G(2d,2p): *m*PW1PW91/6-31G] produces the smallest error for the gas phase NMR chemical shifts.

For the structures with multiple conformers, Boltzmann averaged contribution to the calculated NMR chemical shifts can be used to calculate accurate values. Boltzmann averaged contributions calculated for the various conformers of **2LI1** resulted in a slight improvement in the calculated MAD values of MIM2-calculated NMR chemical shifts. Additionally, in our MIM2-NMR protocol, we found that geometry minimization using molecular mechanics/semi-empirical methods is useful to obtain a good starting geometry to perform the calculations in solution with the implicit, and explicit-implicit solvation models. A closer inspection of the calculated ^{1}H chemical shift revealed that, in most cases, the problematic nuclei are the groups directly bonded

to the amine and amide group. Therefore, we replaced the complex and computationally demanding full explicit solvent box calculations by including a few, directly hydrogen bonded water molecules near the amine and amide groups of the proteins. The bulk solvation effect is then included using the implicit solvation model. With this explicit-implicit solvation model, only one explicit water molecule per amine and amide proton is required to solvate the molecule leading to a significant reduction in the computational cost while maintaining the high-level of accuracy.

Correlation between MIM2-NMR shift predictions and experiment is strong, with the correlation coefficients between 0.98 to 1.0 for ^{13}C and ^{1}H for all the proteins investigated in this paper. With our recommended protocol with MM-restrained minimized structure and explicit-implicit solvation model, a reasonably good accuracy has been achieved: ~0.3 ppm for ^{1}H and ~2.0 ppm for ^{13}C. More importantly, our protocol can be readily applied to structures with the nonstandard residues (i.e., mutations and other functional groups), unlike the empirical treatments such as SHIFTX2 and SHIFTS. The proposed MIM-NMR-explicit-implicit method is accurate and computationally cost-effective and may assist in *de novo* protein structure predictions in the future.

5 Acknowledgments

We acknowledge support from the NSF Grant CHE-1665427 at Indiana University.

References

1. Svergun, D. I.; Petoukhov, M. V.; Koch, M. H. J., Determination of Domain Structure of Proteins from X-Ray Solution Scattering. *Biophys. J.* **2001,** *80*, 2946-2953.

2. Boutet, S.; Lomb, L.; Williams, G. J.; Barends, T. R. M.; Aquila, A.; Doak, R. B.; Weierstall, U.; DePonte, D. P.; Steinbrener, J.; Shoeman, R. L.; Messerschmidt, M.; Barty, A.; White, T. A.; Kassemeyer, S.; Kirian, R. A.; Seibert, M. M.; Montanez, P. A.; Kenney, C.; Herbst, R.; Hart, P.; Pines, J.; Haller, G.; Gruner, S. M.; Philipp, H. T.; Tate, M. W.; Hromalik, M.; Koerner, L. J.; van Bakel, N.; Morse, J.; Ghonsalves, W.; Arnlund, D.; Bogan, M. J.; Caleman, C.; Fromme, R.; Hampton, C. Y.; Hunter, M. S.; Johansson, L. C.; Katona, G.; Kupitz, C.; Liang, M.; Martin, A. V.; Nass, K.; Redecke, L.; Stellato, F.; Timneanu, N.; Wang, D.; Zatsepin, N. A.; Schafer, D.; Defever, J.; Neutze, R.; Fromme, P.; Spence, J. C. H.; Chapman, H. N.; Schlichting, I., High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography. *Science* **2012,** *337*, 362.

3. Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M., Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104*, 9615.

4. Erickson, H. P., Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. *Biol. Proced. Online* **2009,** *11*, 32.

5. Zhou, Z. H., Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **2008,** *18*, 218-228.

6. Morris, G. A., Modern NMR techniques for structure elucidation. *Magn. Reson. Chem.* **1986,** *24*, 371-403.

7. Williams, J. M., Encyclopedia of nuclear magnetic resonance. Volume 1: Historical perspectives. Editors-in-chief D. M. Grant and R. K. Harris. Published by Wiley, Chichester, 1996. ISBN 0-471-95839-5 826 pp. £125, US $195. *Rapid Commun. Mass Spectrom.* **1996,** *10*, 1867-1867.

8. Mulder, F. A. A.; Filatov, M., NMR chemical shift data and ab initio shielding calculations: emerging tools for protein structure determination. *Chem. Soc. Rev.* **2010,** *39*, 578.

9. Helgaker, T.; Jaszuński, M.; Ruud, K., Ab Initio Methods for the Calculation of NMR Shielding and Indirect Spin–Spin Coupling Constants. *Chem. Rev.* **1999,** *99*, 293.

10. Wylie, B. J.; Sperling, L. J.; Nieuwkoop, A. J.; Franks, W. T.; Oldfield, E.; Rienstra, C. M., Ultrahigh resolution protein structures using NMR chemical shift tensors. *Proc. Natl. Acad. Sci. U. S. A.* **2011,** *108*, 16974.

11. Robustelli, P.; Stafford, K. A.; Palmer, A. G., Interpreting Protein Structural Dynamics from NMR Chemical Shifts. *J. Am. Chem. Soc.* **2012,** *134*, 6365-6374.

12. Huang, Y. J.; Brock, K. P.; Ishida, Y.; Swapna, G. V. T.; Inouye, M.; Marks, D. S.; Sander, C.; Montelione, G. T., Combining Evolutionary Covariance and NMR Data for Protein Structure Determination. *Methods Enzymol* **2019,** *614*, 363-392.

13. de Dios, A. C.; Pearson, J. G.; Oldfield, E., Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. *Science* **1993,** *260*, 1491.

14. Casabianca, L. B.; de Dios, A. C., Ab initio calculations of NMR chemical shifts. *J. Chem. Phys.* **2008,** *128*, 052201.

15. He, X.; Wang, B.; Merz, K. M., Protein NMR Chemical Shift Calculations Based on the Automated Fragmentation QM/MM Approach. *J. Phys. Chem. B.* **2009,** *113*, 10380.

16. E., O., CHEMICAL SHIFTS IN AMINO ACIDS, PEPTIDES, AND PROTEINS: From Quantum Chemistry to Drug Design. *Annu. Rev. Phys. Chem.* **2002,** *53*, 349.

17. Willoughby, P. H.; Jansma, M. J.; Hoye, T. R., A guide to small-molecule structure assignment through computation of (1H and 13C) NMR chemical shifts. *Nat. Protoc.* **2014,** *9*, 643.

18. Ballard, C. C.; Hada, M.; Kaneko, H.; Nakatsuji, H., Relativistic study of nuclear magnetic shielding constants: hydrogen halides. *Chem. Phys. Lett.* **1996,** *254*, 170.

19. Kaupp, M.; Bühl, M.; Malkin, V. G., *Calculation of NMR and EPR parameters: theory and applications*. 2004.

20. Vaara, J., Theory and computation of nuclear magnetic resonance parameters. *Phys. Chem. Chem. Phys.* **2007,** *9*, 5399.

21. Facelli, J. C., Calculations of chemical shieldings: Theory and applications. *Concepts Magn. Reson.* **2004,** *20A*, 42.

22. Ditchfield, R., Self-consistent perturbation theory of diamagnetism. *Mol. Phys.* **1974,** *27*, 789.

23. Gauss, J., Effects of electron correlation in the calculation of nuclear magnetic resonance chemical shifts. *J. Chem. Phys.* **1993,** *99*, 3629.

24. Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J., A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J. Chem. Phys.* **1996,** *104*, 5497.

25. Keith, T. A.; Bader, R. F. W., Calculation of magnetic response properties using a continuous set of gauge transformations. *Chem. Phys. Lett.* **1993,** *210*, 223.

26. Keith, T. A.; Bader, R. F. W., Calculation of magnetic response properties using atoms in molecules. *Chem. Phys. Lett.* **1992,** *194*, 1.

27. Hansen, A. E.; Bouman, T. D., Localized orbital/local origin method for calculation and analysis of NMR shieldings. Applications to 13C shielding tensors. *J. Chem. Phys.* **1985,** *82*, 5035.

28. Schindler, M.; Kutzelnigg, W., Theory of magnetic susceptibilities and NMR chemical shifts in terms of localized quantities. II. Application to some simple molecules. *J. Chem. Phys.* **1982,** *76*, 1919.

29. Kutzelnigg, W., Theory of Magnetic Susceptibilities and NMR Chemical Shifts in Terms of Localized Quantities. *Isr. J. Chem.* **1980,** *19*, 193.

30. Rauhut, G.; Puyear, S.; Wolinski, K.; Pulay, P., Comparison of NMR Shieldings Calculated from Hartree–Fock and Density Functional Wave Functions Using Gauge-Including Atomic Orbitals. *J. Phys. Chem.* **1996,** *100*, 6310.

31. Wolinski, K.; Hinton, J. F.; Pulay, P., Efficient implementation of the gauge-independent atomic orbital method for NMR chemical shift calculations. *J. Am. Chem. Soc.* **1990,** *112*, 8251.

32. Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S., SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* **2011,** *50*, 43.

33. Xu, X. P.; Case, D. A., Automated prediction of 15N, 13Cα, 13Cβ and 13C' chemical shifts in proteins using a density functional database. *J. Biomol. NMR* **2001,** *21*, 321.

34. Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M., Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances. *J. Am. Chem. Soc.* **2009,** *131*, 13894-13895.

35. Meiler, J., PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR* **2003,** *26*, 25-37.

36. Swails, J.; Zhu, T.; He, X.; Case, D. A., AFNMR: automated fragmentation quantum mechanical calculation of NMR chemical shifts for biomolecules. *J. Biomol. NMR* **2015,** *63*, 125-139.

37. Lodewyk, M. W.; Siebert, M. R.; Tantillo, D. J., Computational Prediction of 1H and 13C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chem. Rev.* **2012,** *112*, 1839-1862.

38. Hartman, J.; Beran, G., Fragment-based electronic structure approach for computing nuclear magnetic resonance chemical shifts in molecular crystals. *J. Chem. Theory Comput.* **2014,** *10*, 4862.

39. Merz, K. M., Using Quantum Mechanical Approaches to Study Biological Systems. *Acc. Chem. Res.* **2014,** *47*, 2804-2811.

40. Jose, K. V. J.; Raghavachari, K., Fragment-Based Approach for the Evaluation of NMR Chemical Shifts for Large Biomolecules Incorporating the Effects of the Solvent Environment. *J. Chem. Theory Comput.* **2017,** *13*, 1147-1158.

41. Kollwitz, M.; Häser, M.; Gauss, J., Non-Abelian point group symmetry in direct second-order many-body perturbation theory calculations of NMR chemical shifts. *J. Chem. Phys.* **1998,** *108*, 8295.

42. Gauss, J.; Stanton, J. F., Analytic CCSD(T) second derivatives. *Chem. Phys. Lett.* **1997,** *276*, 70.

43. Gauss, J.; Stanton, J. F., Coupled-cluster calculations of nuclear magnetic resonance chemical shifts. *J. Chem. Phys.* **1995,** *103*, 3561.

44. Johnson, B. G.; Frisch, M. J., Analytic second derivatives of the gradient-corrected density functional energy. Effect of quadrature weight derivatives. *Chem. Phys. Lett.* **1993,** *216*, 133.

45. Johnson, B. G.; Fisch, M. J., An implementation of analytic second derivatives of the gradient-corrected density functional energy. *J. Chem. Phys.* **1994,** *100*, 7429.

46. Becke, A. D., Perspective: Fifty years of density-functional theory in chemical physics. *J. Chem. Phys.* **2014,** *140*, 18A301.

47. Jin, X.; Zhu, T.; Zhang, J. Z. H.; He, X., Automated Fragmentation QM/MM Calculation of NMR Chemical Shifts for Protein-Ligand Complexes. *Front. Chem.* **2018,** *6*.

48. Zhu, T.; Zhang, J. Z. H.; He, X., Automated Fragmentation QM/MM Calculation of Amide Proton Chemical Shifts in Proteins with Explicit Solvent Model. *J. Chem. Theory Comput.* **2013,** *9*, 2104-2114.

49. He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. H., Fragment quantum mechanical calculation of proteins and Its applications. *Acc. Chem. Res.* **2014,** *47*, 2748.

50. Sumowski, C. V.; Hanni, M.; Schweizer, S.; Ochsenfeld, C., Sensitivity of ab Initio vs Empirical Methods in Computing Structural Effects on NMR Chemical Shifts for the Example of Peptides. *J. Chem. Theory Comput.* **2014,** *10*, 122-133.

51. Szabó, A.; Ostlund, N. S., *Modern quantum chemistry : introduction to advanced electronic structure theory*. Mineola (N.Y.) : Dover publications: 1996.

52. de Dios, A. C.; Oldfield, E., Methods for computing nuclear magnetic resonance chemical shielding in large systems. Multiple cluster and charge field approaches. *Chem. Phys. Lett.* **1993,** *205*, 108-116.

53. Cui, Q.; Karplus, M., Molecular Properties from Combined QM/MM Methods. 2. Chemical Shifts in Large Molecules. *J. Phys. Chem. B* **2000,** *104*, 3721.

54. Vreven, T.; Morokuma, K.; David, C. S., *Annu. Rep. Comput. Chem.* 2006; Vol. 2, p 35.

55. Chung, L. W.; Hirao, H.; Li, X.; Morokuma, K., The ONIOM method: its foundation and applications to metalloenzymes and photobiology. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012,** *2*, 327.

56. Řezáč, J.; Salahub, D. R., Multilevel Fragment-Based Approach (MFBA): A Novel Hybrid Computational Method for the Study of Large Molecules. *J. Chem. Theory Comput.* **2010,** *6*, 91.

57. Vreven, T.; Byun, K. S.; Komáromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J., Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM. *J. Chem. Theory Comput.* **2006,** *2*, 815.

58. Fedorov, D. G.; Ishida, T.; Kitaura, K., Multilayer Formulation of the Fragment Molecular Orbital Method (FMO). *J. Phys. Chem. A* **2005,** *109*, 2638.

59. Isegawa, M.; Wang, B.; Truhlar, D. G., Electrostatically Embedded Molecular Tailoring Approach and Validation for Peptides. *J. Chem. Theory Comput.* **2013,** *9*, 1381.

60. Beran, G. J. O., Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields. *J. Chem. Phys.* **2009,** *130*, 164115.

61. He, X.; Merz, K. M., Divide and Conquer Hartree–Fock Calculations on Proteins. *J. Chem. Theory Comput.* **2010,** *6*, 405.

62. Nagata, T.; Fedorov, D. G.; Sawada, T.; Kitaura, K.; Gordon, M. S., A combined effective fragment potential–fragment molecular orbital method. II. Analytic gradient and application to the geometry optimization of solvated tetraglycine and chignolin. *J. Chem. Phys.* **2011,** *134*, 034110.

63. Mullin, J. M.; Roskop, L. B.; Pruitt, S. R.; Collins, M. A.; Gordon, M. S., Systematic Fragmentation Method and the Effective Fragment Potential: An Efficient Method for Capturing Molecular Energies. *J. Phys. Chem. A* **2009,** *113*, 10040.

64. Guo, W.; Wu, A.; Zhang, I. Y.; Xu, X., XO: An extended ONIOM method for accurate and efficient modeling of large systems. *J. Comput. Chem.* **2012,** *33*, 2142.

65. Collins, M. A.; Bettens, R. P. A., Energy-Based Molecular Fragmentation Methods. *Chem. Rev.* **2015,** *115*, 5607-5642.

66. Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. A., The Combined Fragmentation and Systematic Molecular Fragmentation Methods. *Acc. Chem. Res.* **2014,** *47*, 2776.

67. Reid, D. M.; Kobayashi, R.; Collins, M. A., Systematic Study of Locally Dense Basis Sets for NMR Shielding Constants. *J. Chem. Theory Comput.* **2014,** *10*, 146.

68. Herbert, J. M., Fantasy versus reality in fragment-based quantum chemistry. *J. Chem. Phys.* **2019,** *151*, 170901.

69. de Dios, A.; Pearson, J.; Oldfield, E., Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. *Science* **1993,** *260*, 1491.

70. Scheurer, C.; Skrynnikov, N. R.; Lienin, S. F.; Straus, S. K.; Brüschweiler, R.; Ernst, R. R., Effects of Dynamics and Environment on 15N Chemical Shielding Anisotropy in Proteins. A Combination of Density Functional Theory, Molecular Dynamics Simulation, and NMR Relaxation. *J. Am. Chem. Soc.* **1999,** *121*, 4242-4251.

71. Exner, T. E.; Frank, A.; Onila, I.; Möller, H. M., Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 3. Conformational Sampling and Explicit Solvents Model. *J. Chem. Theory Comput.* **2012,** *8*, 4818.

72. Gao, Q.; Yokojima, S.; Kohno, T.; Ishida, T.; Fedorov, D. G.; Kitaura, K.; Fujihira, M.; Nakamura, S., Ab initio NMR chemical shift calculations on proteins using fragment molecular orbitals with electrostatic environment. *Chem. Phys. Lett.* **2007,** *445*, 331.

73. Gao, Q.; Yokojima, S.; Fedorov, D. G.; Kitaura, K.; Sakurai, M.; Nakamura, S., Fragment-Molecular-Orbital-Method-Based ab Initio NMR Chemical-Shift Calculations for Large Molecular Systems. *J. Chem. Theory Comput.* **2010,** *6*, 1428.

74. Hartman, J.; Monaco, S.; Schatschneider, B.; Beran, G., Fragment-based 13-C nuclear magnetic resonance chemical shift predictions in molecular crystals: An alternative to planewave methods. *J. Chem. Phys.* **2015,** *143*, 102809.

75. Tan, H. J.; Bettens, R. P. A., Ab initio NMR chemical-shift calculations based on the combined fragmentation method. *Phys. Chem. Chem. Phys.* **2013,** *15*, 7541.

76. Lee, A. M.; Bettens, R. P. A., First Principles NMR Calculations by Fragmentation. *J. Phys. Chem. A* **2007,** *111*, 5111.

77. Zhao, D.; Song, R.; Li, W.; Ma, J.; Dong, H.; Li, S., Accurate Prediction of NMR Chemical Shifts in Macromolecular and Condensed-Phase Systems with the Generalized Energy-Based Fragmentation Method. *J. Chem. Theory Comput.* **2017,** *13*, 5231-5239.

78. Kobayashi, R.; Amos, R. D.; Reid, D. M.; Collins, M. A., Application of the Systematic Molecular Fragmentation by Annihilation Method to ab Initio NMR Chemical Shift Calculations. *J. Phys. Chem. A* **2018,** *122*, 9135-9141.

79. Unzueta, P. A.; Beran, G. J. O., Polarizable continuum models provide an effective electrostatic embedding model for fragment-based chemical shift prediction in challenging systems. *J. Comput. Chem.* **2020,** *41*, 2251-2265.

80. Karadakov, P. B.; Morokuma, K., ONIOM as an efficient tool for calculating NMR chemical shielding constants in large molecules. *Chem. Phys. Lett.* **2000,** *317*, 589.

81. Hall, K. F.; Vreven, T.; Frisch, M. J.; Bearpark, M. J., Three-Layer ONIOM Studies of the Dark State of Rhodopsin: The Protonation State of Glu181. *J. Mol. Biol.* **2008,** *383*, 106.

82. Gascón, J. A.; Sproviero, E. M.; Batista, V. S., QM/MM Study of the NMR Spectroscopy of the Retinyl Chromophore in Visual Rhodopsin. *J. Chem. Theory Comput.* **2005,** *1*, 674.

83. Mayhall, N. J.; Raghavachari, K., Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials. *J. Chem. Theory Comput.* **2011,** *7*, 1336.

84. Jovan Jose, K. V.; Raghavachari, K., Molecules-in-molecules fragment-based method for the evaluation of Raman spectra of large molecules. *Mol. Phys.* **2015,** *113*, 3057.

85. Jovan Jose, K. V.; Raghavachari, K., Raman Optical Activity Spectra for Large Molecules through Molecules-in-Molecules Fragment-Based Approach. *J. Chem. Theory Comput.* **2016,** *12*, 585.

86. Jose, K. V. J.; Beckett, D.; Raghavachari, K., Vibrational Circular Dichroism Spectra for Large Molecules through Molecules-in-Molecules Fragment-Based Approach. *J. Chem. Theory Comput.* **2015,** *11*, 4238-4247.

87. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

88. Thapa, B.; Beckett, D.; Jovan Jose, K. V.; Raghavachari, K., Assessment of Fragmentation Strategies for Large Proteins Using the Multilayer Molecules-in-Molecules Approach. *J. Chem. Theory Comput.* **2018,** *14*, 1383-1394.

89. Thapa, B.; Raghavachari, K., Energy Decomposition Analysis of Protein–Ligand Interactions Using Molecules-in-Molecules Fragmentation-Based Method. *J. Chem. Inf. Model.* **2019,** *59*, 3474-3484.

90. Thapa, B.; Beckett, D.; Erickson, J.; Raghavachari, K., Theoretical Study of Protein–Ligand Interactions Using the Molecules-in-Molecules Fragmentation-Based Method. *J. Chem. Theory Comput.* **2018,** *14*, 5143-5155.

91. Jensen, F., Segmented Contracted Basis Sets Optimized for Nuclear Magnetic Shielding. *Journal of Chemical Theory and Computation* **2015,** *11*, 132-138.

92. Aggelund, P. A.; Sauer, S. P. A.; Jensen, F., Development of polarization consistent basis sets for spin-spin coupling constant calculations for the atoms Li, Be, Na, and Mg. *The Journal of Chemical Physics* **2018,** *149*, 044117.

93. Jensen, F., Basis Set Convergence of Nuclear Magnetic Shielding Constants Calculated by Density Functional Methods. *Journal of Chemical Theory and Computation* **2008,** *4*, 719-727.

94. Cramer, C. J.; Truhlar, D. G., Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999,** *99*, 2161.

95. Pavlíková Přecechtělová, J.; Mládek, A.; Zapletal, V.; Hritz, J., Quantum Chemical Calculations of NMR Chemical Shifts in Phosphorylated Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **2019,** *15*, 5642-5658.

96. Thapa, B.; Raghavachari, K., Accurate pKa Evaluations for Complex Bio-Organic Molecules in Aqueous Media. *J. Chem. Theory Comput.* **2019,** *15*, 6025-6035.

97. Roggatz, C. C.; Lorch, M.; Benoit, D. M., Influence of Solvent Representation on Nuclear Shielding Calculations of Protonation States of Small Biological Molecules. *J. Chem. Theory Comput.* **2018,** *14*, 2684-2695.

98. Semenov, V.; Samultsev, D.; Krivdin, L., Solvent effects in the GIAO-DFT calculations of the 15N NMR chemical shifts of azoles and azines. *Magn. Reson. Chem.* **2014,** *52*.

99. Da Silva, H. C.; De Almeida, W. B., Theoretical calculations of 1H NMR chemical shifts for nitrogenated compounds in chloroform solution. *Chem. Phys.* **2020,** *528*, 110479.

100. Raghavachari, K.; Saha, A., Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.* **2015,** *115*, 5643-5677.

101. Liu, B.; Shadrin, A.; Sheppard, C.; Mekler, V.; Xu, Y.; Severinov, K.; Matthews, S.; Wigneshweraraj, S., A bacteriophage transcription regulator inhibits bacterial transcription initiation by σ-factor displacement. *Nucleic Acids Res.* **2014,** *42*, 4294-4305.

102. Borgert, A.; Heimburg-Molinaro, J.; Song, X.; Lasanajak, Y.; Ju, T.; Liu, M.; Thompson, P.; Ragupathi, G.; Barany, G.; Smith, D. F.; Cummings, R. D.; Live, D., Deciphering Structural Elements of Mucin Glycoprotein Recognition. *ACS Chem. Biol.* **2012,** *7*, 1031-1039.

103. Dračínský, M.; Möller, H. M.; Exner, T. E., Conformational Sampling by Ab Initio Molecular Dynamics Simulations Improves NMR Chemical Shift Predictions. *Journal of Chemical Theory and Computation* **2013,** *9*, 3806-3815.

104. Guerry, P.; Mollica, L.; Blackledge, M., Mapping Protein Conformational Energy Landscapes Using NMR and Molecular Simulation. *ChemPhysChem* **2013,** *14*, 3046-3058.

105. Gerber, P. R.; Müller, K., MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J. Comput.-Aided Mol. Des.* **1995,** *9*, 251-268.

106. Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A., ff14ipq: A Self-Consistent Force Field for Condensed-Phase Simulations of Proteins. *J. Chem. Theory Comput.* **2014,** *10*, 4515-4534.

107. Dahms, S. O.; Könnig, I.; Roeser, D.; Gührs, K.-H.; Mayer, M. C.; Kaden, D.; Multhaup, G.; Than, M. E., Metal Binding Dictates Conformation and Function of the Amyloid Precursor Protein (APP) E2 Domain. *J. Mol. Biol.* **2012,** *416*, 438-452.

108. Dulubova, I.; Ho, A.; Huryeva, I.; Südhof, T. C.; Rizo, J., Three-Dimensional Structure of an Independently Folded Extracellular Domain of Human Amyloid-β Precursor Protein. *Biochemistry* **2004,** *43*, 9583-9588.