JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Automated Fragmentation QM/MM Calculation of Amide Proton Chemical Shifts in Proteins with Explicit Solvent Model

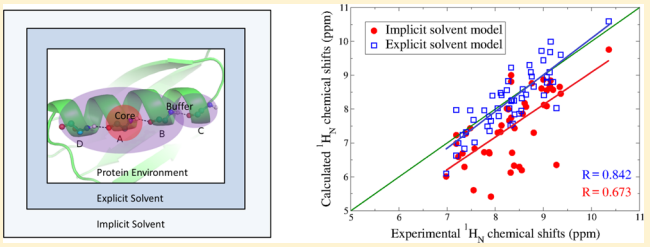
Tong Zhu,[†] John Z. H. Zhang,[†,‡] and Xiao He*,[†]

[†]State Key Laboratory of Precision Spectroscopy and Department of Physics, Institute of Theoretical and Computational Science, East China Normal University, Shanghai, China 200062

[‡]Department of Chemistry, New York University, New York, New York 10003, United States

Supporting Information

ABSTRACT: We have performed a density functional theory (DFT) calculation of the amide proton NMR chemical shift in proteins using a recently developed automated fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) approach. Systematic investigation was carried out to examine the influence of explicit solvent molecules, cooperative hydrogen bonding effects, density functionals, size of the basis sets, and the local geometry of proteins on calculated chemical shifts. Our result demonstrates that the predicted amide proton (^1H_N) NMR chemical shift in explicit solvent shows remarkable improvement over that calculated with the implicit solvation model. The cooperative hydrogen bonding effect is also shown to improve the accuracy of ^1H_N chemical shifts. Furthermore, we found that the OPBE exchange-correlation functional is the best density functional for the prediction of protein ^1H_N chemical shifts among a selective set of DFT methods (namely, B3LYP, B3PW91, M062X, M06L, mPW1PW91, OB98, OPBE), and the locally dense basis set of 6-311++G**/4-31G* is shown to be sufficient for ^1H_N chemical shift calculation. By taking ensemble averaging into account, ^1H_N chemical shifts calculated by the AF-QM/MM approach can be used to validate the performance of various force fields. Our study underscores that the electronic polarization of protein is of critical importance to stabilizing hydrogen bonding, and the AF-QM/MM method is able to describe the local chemical environment in proteins more accurately than most widely used empirical models.



1. INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is one of the most valuable experiment methods used to determine protein structure and investigate dynamics. Great advances have been made in the past decades, resulting in an increase in the precision of NMR measurements and in the size of proteins that can be studied.[1,2] Among the parameters measured by NMR, the chemical shift is an essential output which characterizes the chemical environment of individual atoms.[3−6] Variation of chemical shifts can provide useful information on changes in the local environment of proteins. Recent developments in methods that use chemical shifts to determine high-resolution protein structures help increase the throughput of NMR strategies and promote the role of NMR spectroscopy in structural genomics.[7−14] However, it should be noted that detail of the relationship between the structure and the chemical shifts has not yet been determined well enough for direct structure determination from chemical shifts. To explore the essential factors governing the chemical shifts of protein is still desirable.[15−25]

Over the past decade, quantum chemical (QM) methods have become increasingly useful for NMR chemical shift studies, as they allow one to investigate structural and environmental effects in systematic and controlled manners. Following the pioneering work of de Dios et al.,[26−31] a number of QM studies have been carried out to compute chemical shifts in proteins. Recent methodological advances enabled one to accurately predict chemical shifts in large protein fragments from first principles,[32−35] with the inclusion of important environmental and conformational factors that influence the calculated results. Cui and Karplus proposed a method for calculating chemical shifts in the QM/MM framework[36] and concluded that the QM/MM method can provide good descriptions of the environmental effect on protein NMR chemical shifts. Gao et al. also reported a fragment molecular orbital (FMO) method for NMR chemical shift calculations at the Hartree−Fock (HF) level,[37,38] and their calculations for both α-helix and β-sheet polypeptides agree well with those calculated by conventional HF calculations. Furthermore, Frank et al.[22,23] utilized the adjustable density matrix assembler (ADMA) approach for the QM calculations of protein NMR chemical shifts. The solvent effects on the chemical shifts were also studied by combining them with the PCM method. In our previous studies, a recently developed automated fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) approach was used to calculate NMR chemical shifts for several proteins, and the solvent effects were included by using

Received: November 14, 2012
Published: February 28, 2013

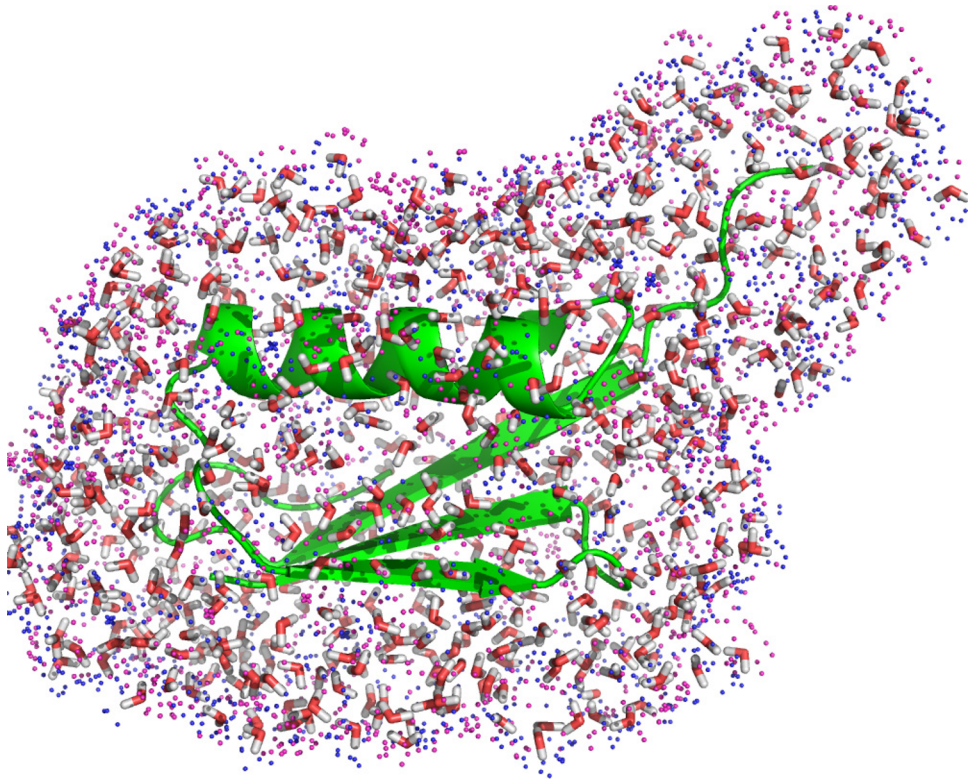


Figure 1. Graphical representation of GB3 (PDB entry: 2IGD) together with the first and second solvation shells and surface charges calculated by DivCon program[46] (red and blue dots represent the positive and negative surface charges, respectively).

the Poisson−Boltzmann (PB) model.[39,40] The AF-QM/MM-PB method has been demonstrated to give results in excellent agreement with that from full quantum calculation, and computed chemical shifts have very good correlation with the experimental values. However, for both the ADMA and AF-QM/MM-PB approaches, the accuracy of predicting the amide proton (^1H_N) chemical shifts of proteins needs to be further improved.

The ^1H_N chemical shift, as one of the most precise NMR parameters that can be measured, plays a key role in peak assignments. Thus, a QM model that can accurately predict the protein ^1H_N chemical shift is in demand. The main reason for the inaccuracy in computed ^1H_N chemical shift arises from the improper treatment of the solvation effect. The specific solvent−solute interactions such as hydrogen bonding cannot be accurately described in implicit solvation models. Explicit inclusion of solvent molecules in the calculation of the ^1H_N chemical shift is required to account for the quantum effect of the solvent. Recently, Exner et al.[41] have introduced the conformational averaging and the explicit water molecules into the ADMA method based on classical molecular dynamic simulations, and a general approach was proposed to calculate protein ^1H, ^{13}C, and ^{15}N chemical shifts. In this study, we focus our attention on accurate calculation of the protein ^1H_N chemical shifts using the AF-QM/MM method. A systematic investigation of the factors that determine the accuracy of the computed ^1H_N chemical shifts was performed. Those factors include the treatment of explicit solvent molecules using a reliable and fast method, cooperative hydrogen bonding effect, the choice of the density functional in DFT calculation, the size of the basis set, and the local geometry of proteins. In addition, the use of ^1H_N chemical shifts calculated by the AF-QM/MM approach to validate the performance of various force fields is also discussed. Our paper is organized as follows: First, the explicit solvent molecules are included in the AF-QM/MM method, and their influence on the calculated ^1H_N chemical shifts is investigated. Second, the influence of the cooperative hydrogen bonding effect, various density functionals, and the size of the basis sets on the accuracy of ^1H_N chemical shift calculation are assessed. Finally, the conformational averaging of the ^1H_N chemical shifts based on snapshots from molecular dynamics (MD) simulation were investigated using two sets of force fields. We also compared the results calculated by the AF-QM/MM-PB method with those derived from several widely used empirical models.

2. COMPUTATIONAL APPROACH

A. AF-QM/MM Method. In the AF-QM/MM approach, the entire protein is divided into nonoverlapping fragments termed core regions. The residues within a certain range from the core region are included in the buffer region. Both the core and the buffer regions are treated by quantum mechanics, while the remainder of the protein is described using an empirical point-charge model to account for the electrostatic effect. Each core-centric (core+buffer) QM/MM calculation is carried out separately, and only the shielding constants of the atoms in the core region are extracted from individual QM/MM calculations. The details of partitioning the system and the definition of the buffer region are described in our previous studies.[39,40] NMR calculation for each fragment is performed with the GIAO method using the Gaussian 09 program.[42] When the calculated ^1H_N chemical shifts are compared with experimental values, the calculated results are referenced to the isotropic shielding constants computed for tetramethylsilane (TMS) at the same level of theory in the gas phase. All the fragment

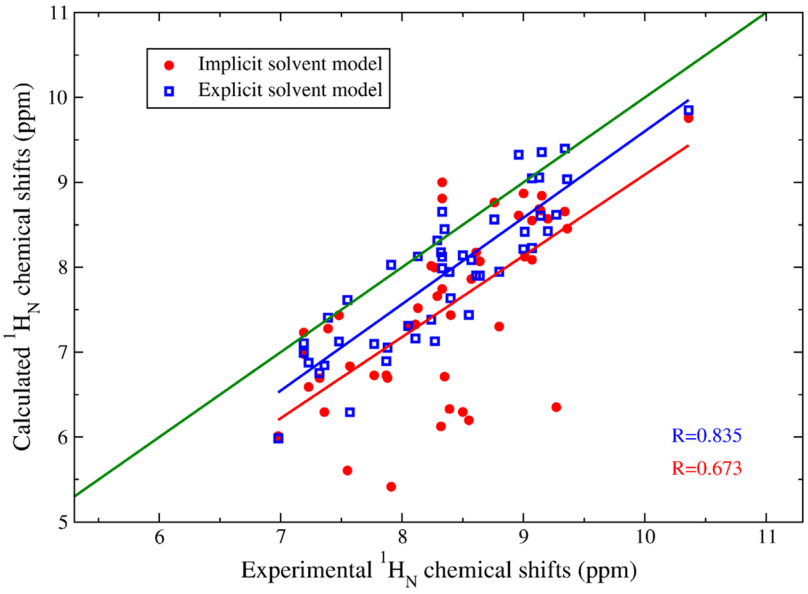


Figure 2. Correlation between experimental and calculated ^1H_N chemical shifts of GB3 using the AF-QM/MM method (the QM level is at B3LYP/6-31G**). Red solid circle, ^1H_N chemical shifts calculated using the implicit solvent model; blue square, ^1H_N chemical shift calculated using the explicit solvent model.

calculations were performed in parallel on a Linux cluster with 8-Core Intel Xeon 2.93 GHz processors.

B. Solvent Model. The main obstacle to including the explicit solvent effect in QM/MM NMR calculation is the determination of positions of solvent molecules around the proteins. Although some crystallographic water molecules are present in X-ray structures, they just represent a small fraction of the water molecules around the proteins. Using standard MD simulation to locate the solvent positions is a formidable task owing to the inefficiency of sampling. A long-time simulation is usually inevitable when trying to converge the flux of solvent and ions throughout the bimolecular interior. In the present study, the distribution of explicit solvent molecules is corrected by the three-dimensional continuous distribution using a 3D reference interaction site model (3D-RISM). The algorithm of the 3D-RISM method is based on statistical mechanics and has been shown to accurately reproduce water distributions at a reduced computational cost. In this work, the PLACEVENT program developed by Hirata and co-workers[43] was utilized to translate the continuous distributions to explicit water molecules. Previous studies have demonstrated that this program places the water molecules on the highest likelihood location and gives excellent agreement with experimental data.[43−45] Only the water molecules in the first and second solvation shells (within 6.0 Å from any atom in the protein) are regarded as part of the entire system. In the AF-QM/MM calculation of protein NMR chemical shift with the explicit solvent model, all the water molecules within 6 Å from any atom of each core region are included in each fragment QM calculation, while all the rest of the water molecules are represented by point charges. The implicit solvent model was used to represent the bulk solvent effect beyond the second solvent shell as shown in Figure 1. To determine the surface charges and protein specific point charges, we use the DivCon program[46,47] which combines the linear-scaling divide-and-conquer semiempirical algorithm with the PB equation to perform the self-consistent reaction field (SCRF) calculation. All the point charges of protein and water molecules are calculated using the PM3/CM2 charge model. The reaction field acting on the solute can be effectively represented by the classical electrostatic potential induced by a set of point charges on the molecular surface as described in our previous study.[39,40]

C. MD Simulation. MD simulations were performed with the Amber ff99SB protein force field using the AMBER11 suite of programs.[48] In the simulation, the protein is placed in a truncated octahedral periodic box of the TIP3P water molecule. The distance from the surface of the box to the closest atom of the solute is set to 12 Å. Counter ions are added to neutralize the system. The entire system is energy minimized using the steepest descent method followed by the conjugate gradient minimization. The system is then heated from 0 to 300 K over 250 ps. Finally, 10 ns MD simulation is performed in the NPT ensemble to further relax the system without any restraints. In MD simulations, the SHAKE algorithm is used to constrain all the bonds involving hydrogen atoms, and the time step is 1 fs. In addition, our previous studies[49−51] have found that MD simulations using the standard Amber force field can break some of the backbone hydrogen bonds in native secondary structures, which results in structural deformation due to a lack of electronic polarization effects. Therefore, in this study, we also employed the recently developed polarized protein-specific charge (PPC)[49] model to provide new atomic charges of protein for a better description of protein dynamics. The PPC charges are fitted at the B3LYP/6-31G** level based on the optimized protein structure using the Amber ff99SB force field. The detailed information of this method can be found in our previous papers.[49,51] To have a direct comparison with the ff99SB force field, a 10 ns simulation also is performed with the PPC model. In the simulation using PPC, the atomic charges of the Amber ff99SB force field are simply replaced by PPC, while the rest of the ff99SB force field parameters are retained.

3. RESULTS AND DISCUSSION

A. Explicit Solvent Model. The X-ray structure of GB3 (PDB entry: 2IGD, 61 residues) which includes both α-helix and β-sheet secondary structures is taken as the initial

Table 1. Comparison of the Experimental and AF-QM/MM Calculated ^1H_N Chemical Shifts (in ppm) of GB3 for Residues Which Form Hydrogen Bonds with Water Molecules Using the Explicit and Implicit Solvent Models, Respectively[a]

residue	LEU12	VAL21	ALA23	GLU24	GLY41	TRP43	THR45
implicit solvation	5.61	6.30	6.13	6.71	5.42	6.35	6.20
explicit solvation	7.62	8.14	8.18	8.45	8.03	8.62	7.94
experiment	7.55	8.50	8.32	8.35	7.91	9.27	8.55

[a]The QM level is at B3LYP/6-31G**.

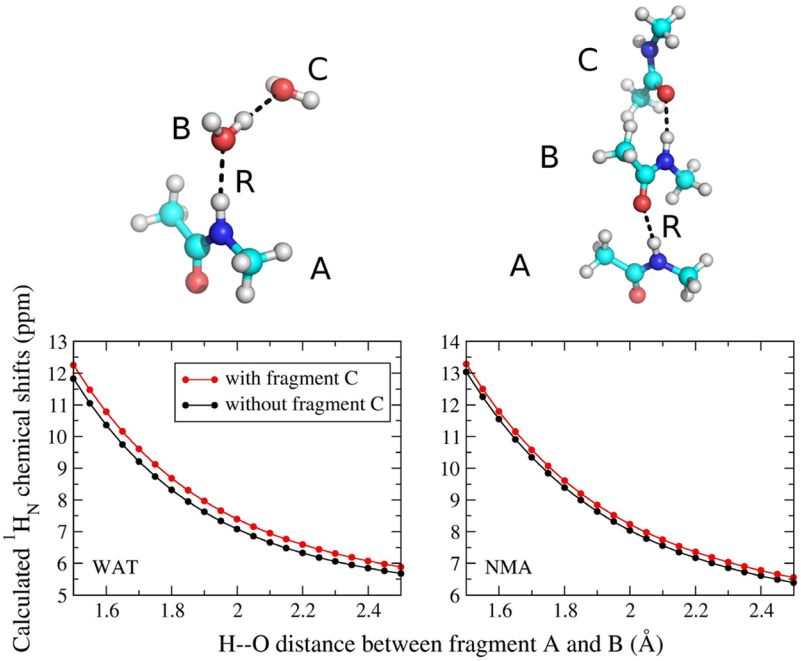


Figure 3. The ^1H_N chemical shift of the central fragment (A) as a function of the ^1H_N−O distance between fragments A and B calculated at the B3LYP/6-311++G** level. Left panel: both the primary and secondary hydrogen bond acceptors are water molecules, Right panel: both the primary and secondary hydrogen bond acceptors are N-methylacetamides (NMAs). The H-bond lengths between fragments B and C are fixed at the original optimized structure at the B3LYP/6-31G** level (1.98 Á for WAT−WAT and 2.09 Á for NMA−NMA, respectively).

geometry. The hydrogen atoms were added by the WHATIF module.[52] Besides the crystallographic water, 678 more water molecules were added by the PLACEVENT program to mimic the first and second solvent shells. A short minimization (1500 steps) was performed to remove the bad contacts in the structure. Calculated ^1H_N chemical shifts using both the explicit and implicit solvent models are compared in Figure 2. To keep consistent with our previous study,[40] the B3LYP/6-31G** method was used. As one can see from Figure 2, the inclusion of explicit water molecules gives considerably better agreement with experimental results over the implicit solvent model. The correlation coefficient (R) between the theoretical and experimental values is improved from 0.673 to 0.835. The root-mean-square error (RMSE) is also decreased from 1.19 to 0.86 ppm. Table 1 lists those residues which have amide protons forming hydrogen bonds (H-bonds) with water molecules. It can be seen from Table 1 that those calculated ^1H_N chemical shifts using the pure implicit solvent model show large upfield shifts as compared to experimental values. When the explicit solvents were included in the fragment QM calculations, the results show significant improvement. It clearly indicates that hydrogen bonding has a large electronic polarization effect on the ^1H_N chemical shift (up to 2−3 ppm). The water molecule which forms direct H-bond with the amide proton in proteins should be treated quantum mechanically to accurately reproduce the experimental ^1H_N chemical shifts.

B. Cooperative Hydrogen Bond Effect. As shown in Figure 2, although the inclusion of explicit water molecules improves the results, the calculated ^1H_N chemical shifts with the explicit solvent model are systematically underestimated by about 0.5 ppm. Previous studies[53−55] on some model systems have illustrated that the cooperative hydrogen bonding effect has a non-negligible influence on ^1H_N chemical shifts by affecting the primary hydrogen bond geometry and polarizing the electron density around the amide proton. In this work, we further explore the cooperative hydrogen bond effect on the protein ^1H_N chemical shifts. For simplicity, we take the N-methylacetamide (NMA) as the central fragment; the cooperative hydrogen bonding effects caused by both water and NMA molecules are investigated. As shown in Figure 3, when the cooperative hydrogen bond was formed, the chemical shifts of the ^1H_N atom in the central residue are downfielded by around 0.3−0.5 ppm as opposed to the case of single H-bond. Therefore, we expand our previous definition of the buffer region (as described in ref 39) to include the secondary hydrogen bond acceptor (the whole residue or water molecule) in the QM region. As depicted in Figure 4, if the ^1H_N chemical shift in the core residue (A) is to be calculated and there is a cooperative hydrogen bond across the peptide bonds of residues A, B (primary H-bond acceptor), and C (secondary

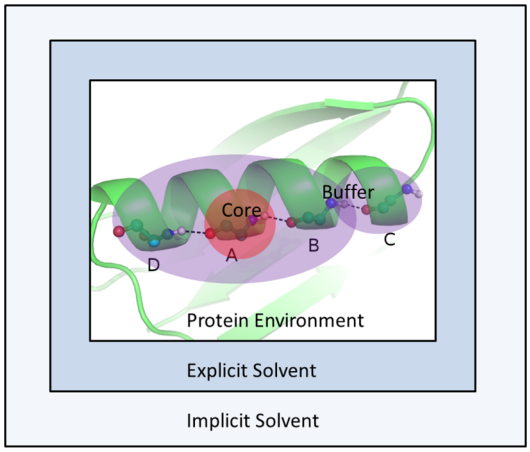


Figure 4. Subsetting scheme for the automated fragmentation AF-QM/MM-PB approach with the explicit solvent model. The red and blue regions represent the core and buffer regions, respectively. On top of the original definition of the buffer region described in ref 39, this study adds one additional criterion which is including the secondary hydrogen bond acceptor (residue C) in the buffer region to take the cooperative hydrogen bonding effect into account. The rest of the protein and explicit solvent molecules are described by point charges. The bulk solvent effect is described by the classical electrostatic potential induced by the point charges on the cavity surface calculated by the PB model.

H-bond acceptor), we also include residue C in the buffer region.

C. The Impact of Density Functional with Locally Dense Basis Set. The other factors that may govern the accuracy of calculated ^1H_N chemical shifts include the density functional and the size of basis set chosen in our calculation. Previous studies on small organic molecules have demonstrated that at least a triple-ζ basis set with the diffuse basis function should be utilized to accurately reproduce the experimental amide hydrogen chemical shift.[56−58] However, the computational cost is very demanding to apply large basis sets on the entire QM region consisting of normally 150−300 atoms, which is the normal size of each fragment (core+buffer region) using the current definition of the buffer region. Therefore, the use of locally dense basis sets, i.e. the combination of two basis sets where the larger one is used for the atoms of interest and the smaller one for all the other atoms, is adopted. The 6-311++G** basis set was employed on the −CO−NH− atoms in both the core residue and other residues involved in the primary and secondary H-bonds (as illustrated in Figure 4). If the H-bond acceptor is a water molecule, the entire water molecule is treated with the 6-311++G** basis set, while the rest of the atoms in the QM region are set to a smaller basis set. In this work, several small basis sets (namely, 3-21G, 4−31G*, 6-31G*, 6-311G**) have been utilized, and their results are compared in Table 2 and Figure 5. As can be seen, the inclusion of the cooperative hydrogen bond effect and applying the locally dense basis set give remarkable improvement for the ^1H_N chemical shifts (compare Figure 5 with Figure 2). Even when a small basis set combination of 6-311++G**/3-21G is employed, the RMSE of ^1H_N chemical shifts is reduced from 0.86 to 0.67 ppm compared to the full B3LYP/6-31G** calculation. The calculation with the B3LYP/6-311++G**/4−31G* method further decreases the RMSE to 0.49 ppm. Moreover, as shown in Table 2, the increase of the basis set from 4-31G* to 6-31G* or 6-311G** does not reduce the

Table 2. Comparison of the Accuracy of AF-QM/MM Calculated ^1H_N Chemical Shifts of GB3 Using Different Mixed Basis Sets with Respect to Experimental Values (MUE: Mean Unsigned Error)

B3LYP/6-311++G**/	RMSE (ppm)	MUE (ppm)	R	correlation function
3-21G	0.67	−0.35	0.831	1.08x-1.13
4-31G*	0.49	−0.06	0.842	1.09x-0.84
6-31G*	0.49	0.04	0.843	1.05x-0.34
6-311G**	0.49	−0.06	0.838	1.00x+0.02

overall RMSE for GB3. Hence, we conclude that the B3LYP functional with the mixed basis set of 6-311++G**/4−31G* strikes a compromise between computational cost and attained accuracy.

We also applied the B3LYP/6-311++G**/4−31G* method to calculate ^1H_N chemical shifts of ubiquitin (PDB entry: 1UBQ, 76 residues). As shown in Figure 6, the correlation with experimental values is similar to that calculated for GB3. Both of them have a correlation of 0.842 (by comparing to the upper right-hand side of Figure 5). It is worth noting that B3LYP/6-311++G**/4−31G* results have better agreement with the experimental values than those using the B3LYP/6-31G** method: the RMSE is 0.53 versus 0.72 ppm and R is 0.842 versus 0.807, indicating that the diffuse basis function is indispensable in capturing the H-bond effect on ^1H_N chemical shift calculations. Furthermore, the calculated ^1H_N chemical shift of residue ILE36 by both the B3LYP/6-311++G**/4−31G* and B3LYP/6-31G** methods (6.11 and 6.13 ppm, respectively) has excellent agreement with the experimental value of 6.15 ppm, whereas the overall agreement of other residues between calculated and experimental results is relatively worse. By examining the local chemical environment of ILE36, we found that the amide proton of ILE36 does not form any hydrogen bond with other residues, and there are no polar groups within 3 Å from the amide proton of ILE36 besides its neighboring residues. This implies that the B3LYP functional may have some deficiencies in describing the local electronic polarization effect of H-bonds on the prediction of ^1H_N NMR chemical shifts.

In previous studies by Vila et al.[57] and Xu and co-workers,[58] various density functionals have been used to calculate NMR chemical shifts of small molecules and relatively large biomolecules. In this study, we further investigate the influence of different density functionals on protein ^1H_N chemical shift calculations. We performed calculations on ^1H_N chemical shifts of GB3 with a selective set of DFT methods (B3LYP, B3PW91, M062X, M06L, mPW1PW91, OB98, OPBE) as compared in Table 3 and Figure 7. In all of the calculations, the mixed basis set of 6-311++G**/4-31G* is used. The reference isotropic shielding constant is computed on the TMS using the corresponding density functional with the 6-311++G** basis set. One can see from Table 3 that the OPBE functional gives the lowest RMSE of 0.47 ppm among all the functionals we have tested here, which is consistent with a previous conclusion by Zhang et al.,[58] where they found that the OPBE functional is one of the best DFT methods for NMR chemical shift calculations. The M062X, B3LYP, and OB98 functionals give a slightly larger RMSE than OPBE, and the B3PW91 functional has the largest RMSE. However, the RMSEs calculated using these seven functionals are very close to each other. We also increased the locally dense basis set to aug-cc-pVTZ or

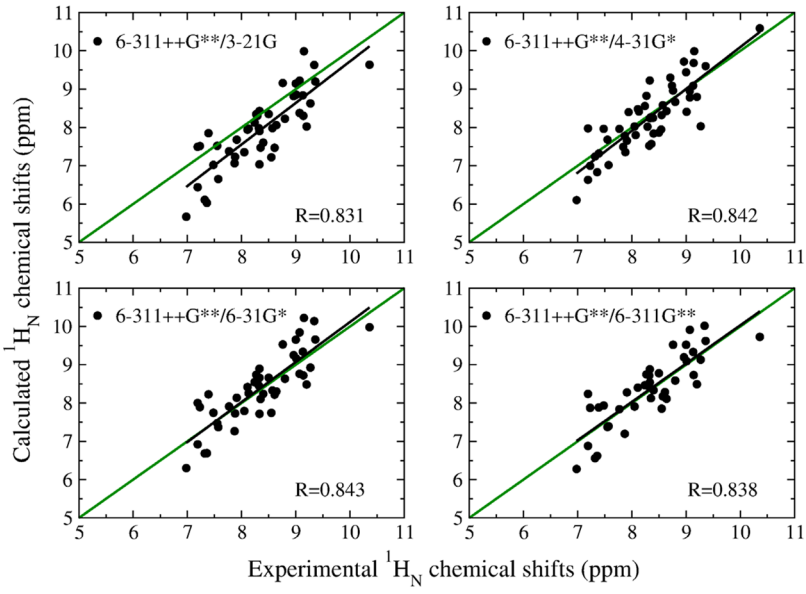


Figure 5. Correlation between the experimental ^1H_N chemical shifts of GB3 and calculated values using the B3LYP functional with various mixed basis sets.

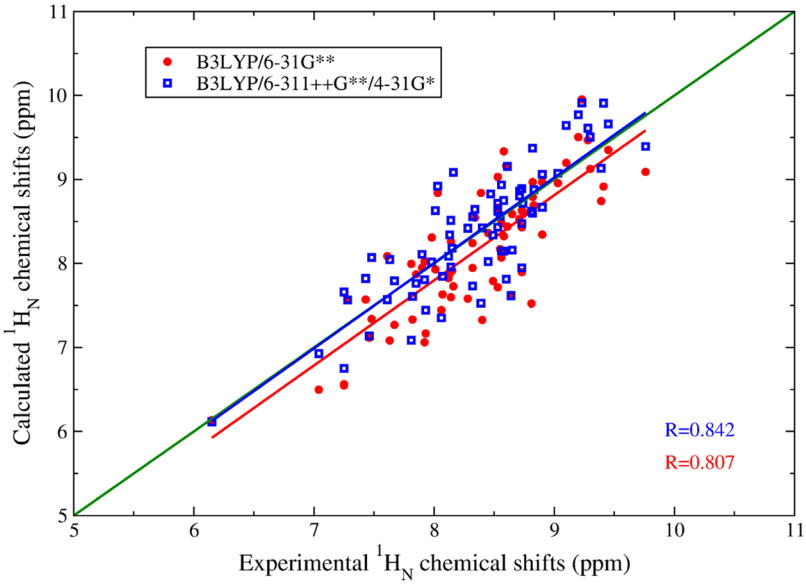


Figure 6. Correlation between the experimental ^1H_N chemical shifts of ubiquitin and calculated values by the AF-QM/MM method using the B3LYP density functional with two different basis sets. The cooperative hydrogen bonding effect is included in both calculations. Red solid circle, B3LYP/6-31G**; blue square, B3LYP/6-311++G**/4-31G*.

Table 3. Performance of AF-QM/MM Calculated ^1H_N Chemical Shifts of GB3 Using Different Density Functionals with the Same Locally Dense Basis Set of 6-311++G/4-31G* with Respect to Experimental Values**

method	RMSE (ppm)	MUE (ppm)	*R*	correlation function
B3LYP	0.49	−0.06	0.842	1.09x-0.84
B3PW91	0.53	−0.06	0.834	1.13x-0.75
M062X	0.48	−0.06	0.845	1.22x-1.34
M06L	0.50	−0.07	0.845	1.20x-1.62
mPW1PW91	0.50	−0.08	0.835	1.12x-1.11
OB98	0.49	−0.05	0.843	1.17x-1.42
OPBE	0.47	−0.05	0.846	1.18x-1.54

expanded the buffer region to include all the residues within 6 Å from any atom of the core residue. As shown in Figure S1 and S2 of the Supporting Information, neither increasing the size of the basis set nor expanding the QM region gives much improvement. As only one structure is used in the calculation, the insufficient conformational sampling may be one of the main causes of the errors; however, the DFT method with current density functionals may also have its inherent limitation in describing the electronic polarization effect of hydrogen bonding for the prediction of ^1H_N NMR chemical shifts.

D. The Influence of Local Geometry in Proteins. So far, all the calculations were performed on a single protein structure. The original X-ray structure 2IGD does not have hydrogen atoms, and thus they were added at idealized positions: a perfect tetrahedral geometry at Cα and the ideal in-

dx.doi.org/10.1021/ct300999w | J. Chem. Theory Comput. 2013, 9, 2104−2114

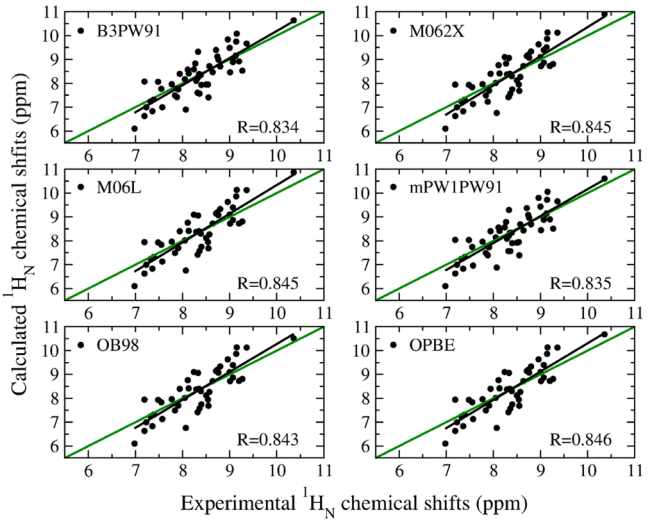


Figure 7. Correlation of ^1H_N chemical shifts of GB3 between the experimental and calculated values using the AF-QM/MM method with the 6-311++G**/4−31G* basis set and various density functionals.

plane position for H_N. In the previous work by Bax and co-workers,[59] NMR measurements of a large set of protein backbone one-bond dipolar couplings have been carried out to refine the structure of GB3 in five different alignment media and consequently produce the accurate positions of H_N atoms relative to the protein backbone. For comparison, we also performed AF-QM/MM calculations at the B3LYP/6-311++G**/4-31G* level on this refined structure by Bax and co-workers (PDB ID: 2OED). As shown in Figure 8, the calculated ^1H_N chemical shifts based on the 2OED structure have better agreement with the experimental values than those calculated using the structure of 2IGD; the RMSE and *R* are 0.45 ppm and 0.877 for 2OED, compared to 0.49 ppm and 0.842 for 2IGD, respectively. However, the hydrogen positions relative to the backbone atoms in a single protein geometry do not have a significant impact on the accuracy of calculated ^1H_N chemical shifts.

E. Ensemble Averaging. Although our AF-QM/MM approach captures all static effects on NMR chemical shifts, it is necessary to take protein dynamics into consideration when we compare calculated NMR chemical shifts directly with experimental values, because any measured chemical shift represents the time- and ensemble-average over fluctuations in protein structure. In this work, we selected 50 snapshots with an interval of 100 ps from the last 5 ns of the MD simulation trajectory after the system was well equilibrated using both the Amber ff99SB force field and PPC charge model. For each structure, ^1H_N chemical shifts were calculated by the AF-QM/MM method at the B3LYP/6-311++G**/4-31G* level. Figure 9 shows the correlation between the experimental and ensemble-averaged ^1H_N chemical shifts of GB3 for the two charge models over 50 snapshots. The ensemble averaged result from the PPC model shows significant improvement over that using the Amber ff99SB charge model. The RMSEs of ^1H_N chemical shifts are 0.63 and 1.27 ppm for the PPC and ff99SB charge model, respectively. However, owing to a small sampling set and the limitation of the empirical potentials, the ensemble averaged result from the PPC model is slightly worse than that calculated on the single crystal structure. We further selected two residues from GB3: one is located on the solvent accessible surface (TRP 43); the other one is buried deeply inside the protein (ASP 46). The probability distribution of these two ^1H_N chemical shifts is calculated from 500 snapshots. As shown in Figure S3 of the Supporting Information, the distribution of calculated ^1H_N chemical shifts is well correlated with the distribution of the hydrogen bond length, and the PPC charge model gives better agreement with experimental observations. Exner et al.[41] have concluded that thousands of snapshots are needed to get a reasonable distribution of the calculated chemical shifts. Owing to significant computational cost of QM calculations on large proteins with thousands of configurations, performing single point calculation on the X-ray structure of protein with explicit water molecules added by the 3D-RISM

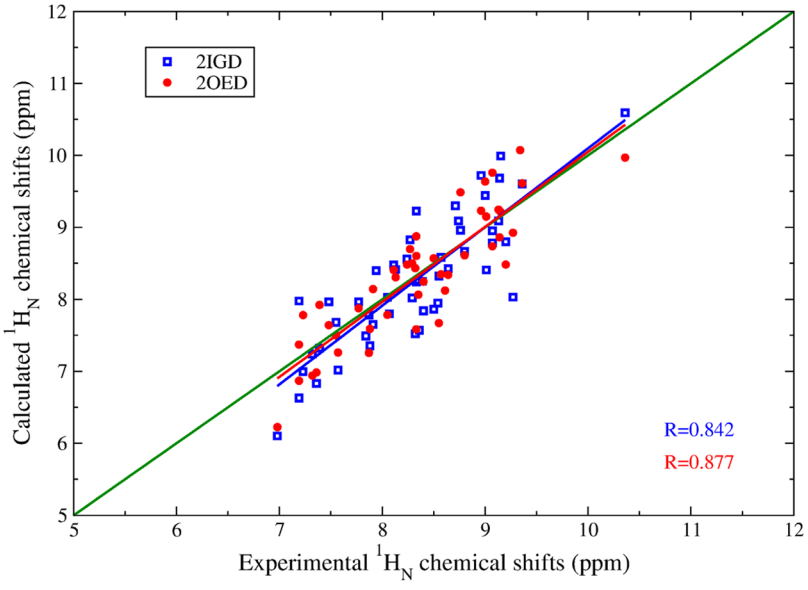


Figure 8. Correlation between the experimental and calculated ^1H_N chemical shifts of GB3 on two different PDB structures using the AF-QM/MM method at the B3LYP/6-311++G**/4−31G* level (blue square, PDB entry 2IGD; red solid circle, PDB entry 2OED).

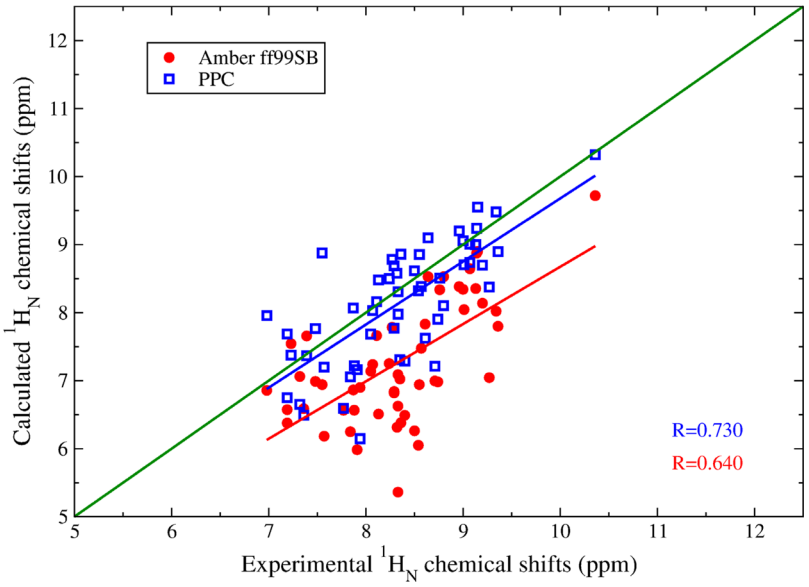


Figure 9. Correlation between the experimental and ensemble-averaged ^1H_N chemical shifts of GB3. Red solid circle: ensemble averaging over 50 snapshots taken from the MD simulated trajectory using the Amber ff99SB force field. Blue square: ensemble averaging over 50 snapshots taken from the MD simulated trajectory using the PPC charge model.

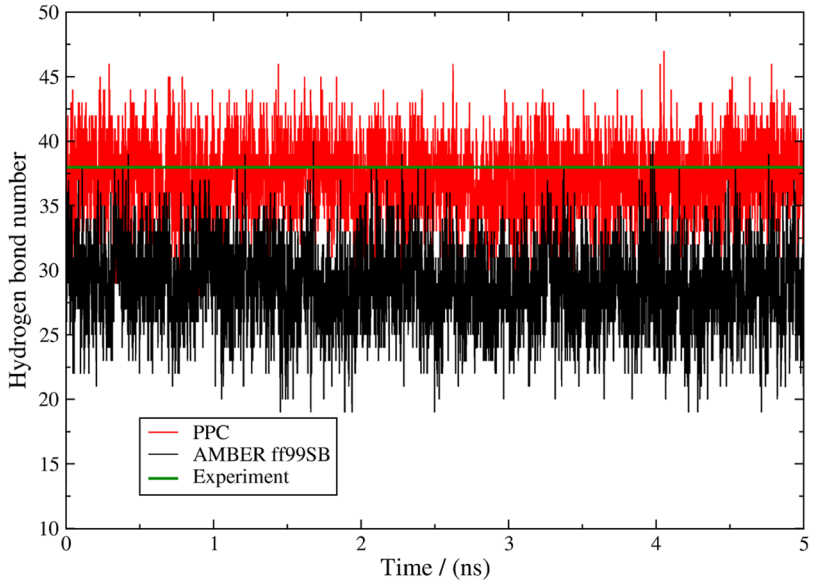


Figure 10. Comparison of the number of H-bonds as a function of MD simulation time using the Amber ff99SB force field (black) and the PPC charge model (red), respectively. The number of H-bonds in the crystal structure is shown in green. The H-bond is defined based on the following criteria: the distance between N and O is less than 3.2 Å, and the angle of N−H···O is greater than 110°.

method may be a fast and reliable approach to predict the protein ^1H_N chemical shifts.

As the ^1H_N chemical shifts are very sensitive to the local H-bond geometry, the failure of the ff99SB force field mainly originates from the breaking or deformation of some intraprotein or protein−solvent hydrogen bonds due to a lack of the electronic polarization effect. We further plot the number of H-bonds as a function of MD simulation time as shown in Figure 10. One can see from the figure that more H-bonds are preserved in the PPC simulation than in the ff99SB simulation, and the time-averaged number of H-bonds from PPC simulation is very close to the experimental value of 38 (including both the intraprotein and protein−solvent H-bonds). The result clearly shows that the electronic polarization is important in stabilizing the H-bonds, which is critical to preserving the native secondary structure of proteins. The calculated ^1H_N chemical shifts accurately reflect the deformation of H-bonds across the structure during MD simulation using the nonpolarizable ff99SB force field. In addition, it indicates that the presence of the water molecules inside the protein or the bridging water between the ligand and protein can be determined by comparing the calculated ^1H_N chemical shifts directly with experimental values. As the protein chemical shifts can be precisely measured and are widely accessible, they have been widely used to validate the empirical force field.[60,61] Most previous studies focused on the C_α and C_β atoms, which can reflect the accuracy of the backbone torsions. Our study demonstrates that the ^1H_N chemical shifts can also be utilized

to test the performance of force fields, especially for the strength of hydrogen bond and other electronic polarization effects.

F. Comparison with Empirical Models. Finally, we compare the calculated results of ^1H_N chemical shifts of GB3 and ubiquitin using the AF-QM/MM method at the B3LYP/6-311++G**/4-31G* level with those calculated using four popular empirical models: SHIFTX2,[62] SHIFTS,[63] SPARTA+,[64] and CAMSHIFT.[65] The results are compared in Table 4 and Figure 11. Clearly, SHIFTX2 gives the best result among these empirical methods. The RMSE and R calculated by SHIFTX2 are 0.28 ppm and 0.927, respectively. On the other hand, the RMSE from the AF-QM/MM calculation is 0.59 ppm, which is lower than that calculated by SHIFTS, but larger than the other three empirical methods: SHIFTX2, SPARTA+, and CAMSHIFT. However, it should be noted that the slope of the correlation function between AF-QM/MM results and experiment is almost exactly 1. In contrast, the slopes given by all the empirical methods are much smaller than 1. As shown in Figure 11, ^1H_N chemical shifts calculated by all the empirical methods fall into a narrower range than experimental results, indicating that the ^1H_N chemical shifts are overly fitted through the parametrization of the empirical models. Furthermore, for the ^1H_N chemical shift of residue ILE36, which does not form any hydrogen bond with other residues in ubiquitin, all the results calculated by empirical methods are significantly overestimated as compared to the experimental value (see Table 5). On the contrary, the AF-QM/MM method gives excellent agreement with the experiment. Since the AF-QM/MM method treats the nearby residues quantum mechanically, it can reflect the local chemical environment in proteins more precisely than empirical models. In addition, the AF-QM/MM methods can be readily extended to more general biological systems than current empirical methods, such as nonstandard residues, metalloproteins, protein−ligand, protein−DNA/RNA, and membrane protein−lipid complexes. Research along these directions is underway in our laboratory.

Table 4. Comparison of the Calculated ^1H_N Chemical Shifts of GB3 and Ubiquitin Using the AF-QM/MM Method at the B3LYP/6-311++G/4-31G* Level and Four Empirical Models with Respect to Experimental Values**

method	RMSE (ppm)	MUE (ppm)	R	correlation function
AF-QM/MM	0.59	−0.06	0.793	1.00x-0.08
SHIFTX2	0.28	0.03	0.927	0.82x+1.51
SHIFTS	0.64	−0.10	0.577	0.45x+4.57
SPARTA+	0.42	0.04	0.833	0.67x+2.82
CAMSHIFT	0.46	−0.04	0.752	0.59x+3.38

Table 5. Comparison of ^1H_N Chemical Shifts of Residue ILE36 in Ubiquitin Calculated with the AF-QM/MM Method at the B3LYP/6-311++G/4-31G* Level and Four Empirical Models**

exptl.	AF-QM/MM	SHIFTX2	SHIFTS	SPARTA+	CAMSHIFT
6.15	6.11	6.80	7.75	7.59	7.23

4. CONCLUSIONS

We have performed DFT calculations of protein amide proton NMR chemical shifts using our recently developed AF-QM/MM approach. The solvent effects were included by treating the explicit water molecules in the first and second solvation shells quantum mechanically. The calculated ^1H_N chemical shift using the explicit solvent model shows remarkable improvement over that from the implicit solvent calculation. The cooperative hydrogen bonding effect is also shown to improve the accuracy of ^1H_N chemical shifts. Moreover, we assessed the performance of different density functionals and the size of the basis set on the protein ^1H_N chemical shift calculations. The results demonstrate that the OPBE exchange-correlation functional with the locally dense basis set of 6-311++G**/4-31G* is one of the best density functionals and mixed basis sets

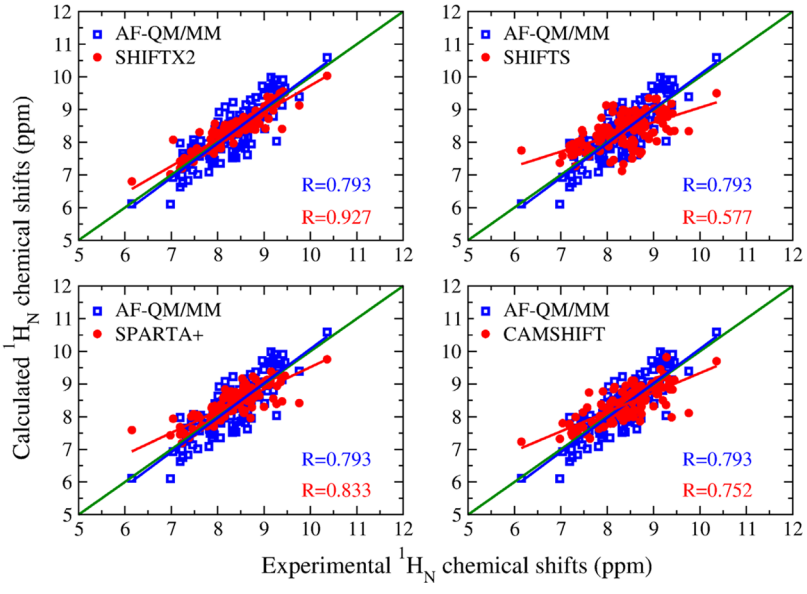


Figure 11. Correlation of ^1H_N chemical shifts of GB3 and ubiquitin between the experimental and calculated values using the AF-QM/MM method at the B3LYP/6-311++G**/4-31G* level (blue squares) and four empirical methods: SHIFTX2, SHIFTS, SPARTA+, and CAMSHIFT (red solid circles).

for the prediction of protein ^1H_N chemical shift. However, the density functional still needs to be improved to achieve higher accuracy on calculating the protein ^1H_N chemical shifts.

The ensemble averaging of the ^1H_N chemical shift is also calculated over the snapshots from the MD simulation. The ^1H_N chemical shift can be used as a probe to test the stability of intraprotein hydrogen bonding and further certify the presence of water molecules inside the protein structure or the bridging water between the ligand and protein. Our study underscores that the electronic polarization is critical to stabilizing the H-bonds in proteins. The ^1H_N chemical shifts calculated by the AF-QM/MM method can be utilized as a benchmark test to evaluate the accuracy of molecular force fields in describing the H-bond strength during the MD simulation.

Finally, we compared the AF-QM/MM calculated ^1H_N chemical shifts with other empirical models. The results show that the overall RMSE of the AF-QM/MM result is comparable to most of the empirical methods; even the empirical models are overly fitted against experimental data. What's more, the AF-QM/MM method is able to describe the local chemical environment in proteins more accurately than empirical methods. The AF-QM/MM approach with the explicit solvation treatment is computationally efficient and linear-scaling with a low prefactor. The approach is massively parallel and can be applied to routinely calculate the *ab initio* NMR chemical shifts for proteins of any size. The applications extended to more general biological systems are under investigation in our laboratory.

■ ASSOCIATED CONTENT

Supporting Information

Several figures providing additional data and information from our calculations for further detailed analysis and comparison are provided. This information is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: xiaohe@phy.ecnu.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 10974054 and 20933002) and Shanghai PuJiang program (09PJ1404000). We also thank the Computational Center of ECNU for providing us computational time.

■ REFERENCES

(1) Bieri, M.; Kwan, A. H.; Mobli, M.; King, G. F.; Mackay, J. P.; Gooley, P. R. *FEBS J.* **2011**, *278*, 704−715.
(2) Kwan, A. H.; Mobli, M.; Gooley, P. R.; King, G. F.; Mackay, J. P. *FEBS J.* **2011**, *278*, 687−703.
(3) Saito, H.; Ando, I.; Ramamoorthy, A. *Prog. Nucl. Magn. Reson. Spectrosc.* **2010**, *57*, 181−228.
(4) Baskaran, K.; Brunner, K.; Munte, C. E.; Kalbitzer, H. R. *J. Biomol. NMR* **2010**, *48*, 71−83.
(5) Schumann, F. H.; Riepl, H.; Maurer, T.; Gronwald, W.; Neidig, K.-P.; Kalbitzer, H. R. *J. Biomol. NMR* **2007**, *39*, 275−289.
(6) Oldfield, E. *Annu. Rev. Phys. Chem.* **2002**, *53*, 349−378.
(7) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *P. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 9615−9620.
(8) Tolbert, B. S.; Miyazaki, Y.; Barton, S.; Kinde, B.; Starck, P.; Singh, R.; Bax, A.; Case, D. A.; Summers, M. F. *J. Biomol. NMR* **2010**, *47*, 205−219.
(9) Sahakyan, A. B.; Vranken, W. F.; Cavalli, A.; Vendruscolo, M. *Angew. Chem., Int. Ed.* **2011**, *50*, 9620−9623.
(10) Avbelj, F.; Kocjan, D.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17394−17397.
(11) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 4685−4690.
(12) Wylie, B. J.; Sperling, L. J.; Nieuwkoop, A. J.; Franks, W. T.; Oldfield, E.; Rienstra, C. M. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 16974−17979.
(13) Stratmann, D.; Boelens, R.; Bonvin, A. M. J. J. *Proteins* **2011**, *79*, 2662−2670.
(14) Kutzelnigg, W. *J. Mol. Struct: THEOCHEM* **1989**, *202*, 11−61.
(15) Wishart, D. S. *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *58*, 62−87.
(16) Mulder, F. A. A.; Filatov, M. *Chem. Soc. Rev.* **2010**, *39*, 578−590.
(17) Wilton, D. J.; Kitahara, R.; Akasaka, K.; Williamson, M. P. *J. Biomol. NMR* **2009**, *44*, 25−33.
(18) Tomlinson, J. H.; Green, V. L.; Baker, P. J.; Williamson, M. P. *Proteins* **2010**, *78*, 3000−3016.
(19) Cioffi, M.; Hunter, C. A.; Packer, M. J.; Pandya, M. J.; Williamson, M. P. *J. Biomol. NMR* **2009**, *43*, 11−19.
(20) Yao, L.; Grishaev, A.; Cornilescu, G.; Bax, A. *J. Am. Chem. Soc.* **2010**, *132*, 10866−10875.
(21) Moon, S.; Case, D. A. *J. Biomol. NMR* **2007**, *38*, 139−150.
(22) Frank, A.; Onila, I.; Möller, H. M.; Exner, T. E. *Proteins* **2011**, *79*, 2189−2202.
(23) Frank, A.; Möller, H. M.; Exner, T. E. *J. Chem. Theory Comput.* **2012**, *8*, 1480−1492.
(24) Moon, S.; Case, D. A. *J. Comput. Chem.* **2006**, *27*, 825−836.
(25) Tang, S.; Case, D. A. *J. Biomol. NMR.* **2011**, *51*, 303−312.
(26) de Dios, A. C.; Oldfield, E. *Chem. Phys. Lett.* **1993**, *205*, 108−116.
(27) de Dios, A. C. *Prog. Nucl. Magn. Reson. Spectrosc.* **1996**, *29*, 229−278.
(28) Facelli, J. C.; de Dios, A. C. *Modeling NMR Chemical Shifts: Gaining Insights into Structure and Environment*; Oxford University Press: Oxford, U. K., 1999.
(29) de Dios, A. C.; Pearson, J. G.; Oldfield, E. *Science* **1993**, *260*, 1491−1496.
(30) Wang, B.; Brothers, E. N.; van der Vaart, A.; Merz, K. M. *J. Chem. Phys.* **2004**, *120*, 11392−11400.
(31) Vila, J. A.; Aramini, J. M.; Rossi, P.; Kuzin, A.; Su, M.; Seetharaman, J.; Xiao, R.; Tong, L.; Montelione, G. T.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 14389−14394.
(32) van Mourik, T. *J. Chem. Phys.* **2006**, *125*, 191101−191104.
(33) Dracinsky, M.; Bour, P. *J. Chem. Theory Comput.* **2010**, *6*, 288−299.
(34) Flaig, D.; Beer, M.; Ochsenfeld, C. *J. Chem. Theory Comput.* **2012**, *8*, 2260−2271.
(35) Beer, M.; Kussmann, J.; Ochsenfeld, C. *J. Chem. Phys.* **2011**, *134*, 74102−74116.
(36) Cui, Q.; Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 3721−3743.
(37) Gao, Q.; Yokojima, S.; Kohno, T.; Ishida, T.; Fedorov, D. G.; Kitaura, K.; Fujihira, M.; Nakamura, S. *Chem. Phys. Lett.* **2007**, *445*, 331−339.
(38) Gao, Q.; Yokojima, S.; Fedorov, D. G.; Kitaura, K.; Sakurai, M.; Nakamura, S. *J. Chem. Theory Comput.* **2010**, *6*, 1428−1444.
(39) He, X.; Wang, B.; Merz, K. M. *J. Phys. Chem. B.* **2009**, *113*, 10380−10388.
(40) Zhu, T.; He, X.; Zhang, J. Z. H. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7837−7845.
(41) Exner, T. E.; Frank, A.; Onila, I.; Möller, H. M. *J. Chem. Theory Comput.* **2012**, *8*, 4818−4827.

(42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision B.01; Gaussian, Inc.: Wallingford, CT, 2010.
(43) Sindhikara, D. J.; Yoshida, N.; Hirata, F. *J. Comput. Chem.* **2012**, *33*, 1536−1543.
(44) Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. *Proteins* **2007**, *66*, 804−813.
(45) Yoshida, N.; Phongphanphanee, S.; Maruyama, Y.; Imai, T.; Hirata, F. *J. Am. Chem. Soc.* **2006**, *128*, 12042−12043.
(46) Dixon, S. L.; van der Vaart, A.; Gogonea, V.; Vincent, M.; Brothers, E. N.; Suarez, D.; Westerhoff, L. M.; Merz, K. M., Jr. *DivCon*; The Pennsylvania State University: University Park, PA, 1999.
(47) Gogonea, V.; Merz, K. M. *J. Phys. Chem. A* **1999**, *103*, 5171−5188.
(48) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668−1688.
(49) Ji, C. G.; Zhang, J. Z. H. *J. Phys. Chem. B.* **2009**, *113*, 13898−13900.
(50) Ji, C. G.; Zhang, J. Z. H. *J. Phys. Chem B* **2009**, *113*, 16059−16064.
(51) Duan, L. L.; Mei, Y.; Zhang, Q. G.; Zhang, J. Z. H. *J. Chem. Phys.* **2009**, *130*, 115102−115107.
(52) Vriend, G. *J. Mol. Graphics* **1990**, *8*, 52−56.
(53) Barfield, M. *J. Am. Chem. Soc.* **2002**, *124*, 4158−4168.
(54) Jiang, X. N.; Wang, C. S. *ChemPhysChem* **2009**, *10*, 3330−3336.
(55) Parker, L. L.; Houk, A. R.; Jensen, J. H. *J. Am. Chem. Soc.* **2006**, *128*, 9863−9872.
(56) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293−352.
(57) Vila, J. A.; Baldoni, H. A.; Scheraga, H. A. *J. Comput. Chem.* **2009**, *30*, 884−892.
(58) Zhang, Y.; Wu, A.; Xu, X.; Yan, Y. *Chem. Phys. Lett.* **2006**, *421*, 383−388.
(59) Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 9179−9191.
(60) Li, D. W.; Brüschweiler, R. *Angew. Chem., Int. Ed.* **2010**, *49*, 6778−6780.
(61) Li, D. W.; Brüschweiler, R. *J. Phys. Chem. Lett.* **2010**, *1*, 246−248.
(62) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. *J. Biomol. NMR* **2011**, *50*, 43−57.
(63) Xu, X. P.; Case, D. A. *J. Biomol. NMR* **2001**, *21*, 321−333.
(64) Shen, Y.; Bax, A. *J. Biomol. NMR* **2010**, *48*, 13−22.
(65) Sahakyan, A. B.; Vranken, W. F.; Cavalli, A.; Vendruscolo, M. *J. Biomol. NMR* **2011**, *50*, 331−346.