ACCOUNTS of chemical research

Article

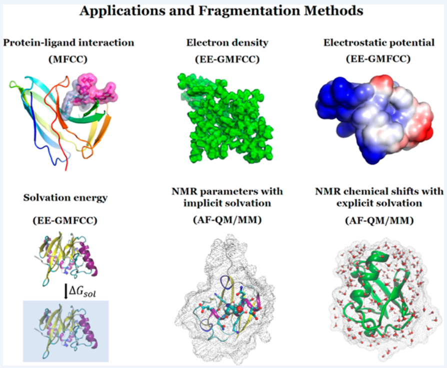
pubs.acs.org/accounts

Fragment Quantum Mechanical Calculation of Proteins and Its Applications

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

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

[‡]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China


Applications and Fragmentation Methods
Protein-ligand interaction (MFCC)
Electron density (EE-GMFCC)
Electrostatic potential (EE-GMFCC)
Solvation energy (EE-GMFCC)
ΔG_{sol}
NMR parameters with implicit solvation (AF-QM/MM)
NMR chemical shifts with explicit solvation (AF-QM/MM)

CONSPECTUS: The desire to study molecular systems that are much larger than what the current state-of-the-art *ab initio* or density functional theory methods could handle has naturally led to the development of novel approximate methods, including semiempirical approaches, reduced-scaling methods, and fragmentation methods. The major computational limitation of *ab initio* methods is the scaling problem, because the cost of *ab initio* calculation scales *n*th power or worse with system size. In the past decade, the fragmentation approach based on chemical locality has opened a new door for developing linear-scaling quantum mechanical (QM) methods for large systems and for applications to large molecular systems such as biomolecules. The fragmentation approach is highly attractive from a computational standpoint. First, the *ab initio* calculation of individual fragments can be conducted almost independently, which makes it suitable for massively parallel computations. Second, the electron properties, such as density and energy, are typically combined in a linear fashion to reproduce those for the entire molecular system, which makes the overall computation scale linearly with the size of the system.

In this Account, two fragmentation methods and their applications to macromolecules are described. They are the electrostatically embedded generalized molecular fractionation with conjugate caps (EE-GMFCC) method and the automated fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) approach. The EE-GMFCC method is developed from the MFCC approach, which was initially used to obtain accurate protein–ligand QM interaction energies. The main idea of the MFCC approach is that a pair of conjugate caps (concaps) is inserted at the location where the subsystem is divided by cutting the chemical bond. In addition, the pair of concaps is fused to form molecular species such that the overcounted effect from added concaps can be properly removed. By introducing the electrostatic embedding field in each fragment calculation and two-body interaction energy correction on top of the MFCC approach, the EE-GMFCC method is capable of accurately reproducing the QM molecular properties (such as the dipole moment, electron density, and electrostatic potential), the total energy, and the electrostatic solvation energy from full system calculations for proteins.

On the other hand, the AF-QM/MM method was used for the efficient QM calculation of protein nuclear magnetic resonance (NMR) parameters, including the chemical shift, chemical shift anisotropy tensor, and spin–spin coupling constant. In the AF-QM/MM approach, each amino acid and all the residues in its vicinity are automatically assigned as the QM region through a distance cutoff for each residue-centric QM/MM calculation. Local chemical properties of the central residue can be obtained from individual QM/MM calculations. The AF-QM/MM approach precisely reproduces the NMR chemical shifts of proteins in the gas phase from full system QM calculations. Furthermore, via the incorporation of implicit and explicit solvent models, the protein NMR chemical shifts calculated by the AF-QM/MM method are in excellent agreement with experimental values. The applications of the AF-QM/MM method may also be extended to more general biological systems such as DNA/RNA and protein–ligand complexes.

1. INTRODUCTION

At present, accurate and efficient quantum chemistry calculations for macromolecules (containing more than 500 atoms) still present a grand challenge to computational chemists. The major limitation of *ab initio* methods is the scaling problem.[1] At the Hartree–Fock (HF) and density functional theory (DFT) levels, the conventional high-power scaling is $O(N^3)$ (N denotes the size of the system). As for post-HF methods, second-order Møller–Plesset perturbation theory (MP2) scales as $O(N^5)$, and the coupled-cluster (CC) method that includes single and double excitations (CCSD) scales as $O(N^6)$.

Much effort has been devoted to the development of linear-scaling methods for energy calculation of large molecular systems at the *ab initio* level.[2−10] In recent years, fragmentation

Special Issue: Beyond QM/MM: Fragment Quantum Mechanical Methods

Received: February 25, 2014
Published: May 22, 2014

dx.doi.org/10.1021/ar500077t | Acc. Chem. Res. 2014, 47, 2748−2757

methods emerged as highly efficient and powerful approaches for developing linear-scaling QM methods for large systems.[11] The fragmentation approach is based on the "chemical locality" of macromolecular systems, which assumes that the local region of a macromolecule is only weakly influenced by the atoms that are far from the region of interest. On the basis of this chemical intuition, the macromolecule is usually divided into subsystems (fragments) in the fragmentation approaches, and subsequently, the total energy or molecular properties of the whole system can be obtained by taking a proper linear combination of the corresponding terms of individual fragments.

The fragmentation approach is more attractive as a practical tool for electronic structure calculation of large systems than conventional linear-scaling methods in several aspects, such as easy implementation of parallelization without extensively modifying the existing QM programs and straightforward application at all levels of *ab initio* electronic structure theories. Over the past decade, a range of fragmentation QM methods for large systems have been proposed, including the fragment molecular orbital (FMO) method,[12−15] the molecular fractionation with conjugate caps (MFCC) approach,[16−21] the systematic fragmentation method (SFM),[22−24] the adjustable density matrix assembler (ADMA) approach,[25−27] the molecular tailoring approach (MTA),[28−30] the generalized energy-based fragmentation (GEBF) method,[31−33] the electrostatically embedded many-body (EE-MB) expansion approach,[34,35] the explicit polarization (X-Pol) potential,[36,37] and the automated fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) method.[38−40] A comprehensive review of the fragmentation QM methods can be found in a recent review by Gordon and co-workers.[11]

The fragmentation methods have been successfully applied in various applications to complex molecular systems such as molecular clusters, proteins, and protein−ligand complexes. In this Account, we describe two fragmentation QM methods and their applications. One is the electrostatically embedded generalized MFCC (EE-GMFCC) method,[41,42] which is a more generalized approach to computing the total energy and molecular properties of macromolecules on top of the MFCC method. The other is the AF-QM/MM method for the calculation of protein nuclear magnetic resonance (NMR) chemical shifts by combination with implicit and explicit solvation models.[38−40,43−45]

2. MFCC METHOD

2.1. Protein−Ligand Interaction Energy

The MFCC method was initially aimed to provide efficient, linear-scaling *ab initio* calculation of protein−ligand interaction energies.[16] The main idea of the MFCC approach is to divide a protein molecule into amino acid fragments that are properly capped.[16,46−48] Using the fractionation scheme, the interaction energy between the protein and ligand can be computed by separate calculations of individual fragments interacting with the ligand. A crucial feature of the MFCC approach is that a pair of conjugate caps (concaps) is inserted at the cutting location. These caps are introduced to serve two purposes. (1) They cap the cutoff fragments to saturate the dangling bonds. (2) They mimic the local chemical environment of the original protein to the cutoff fragments. In addition, the pair of concaps is fused to form proper molecular species such that the doubly counted interaction energy between caps of the fragments and the ligand can be thoroughly subtracted. Hydrogen atoms are added to terminate the molecular caps to prevent dangling bonds. Figure 1 illustrates the MFCC scheme in which a peptide bond is cut and the fragments are capped.

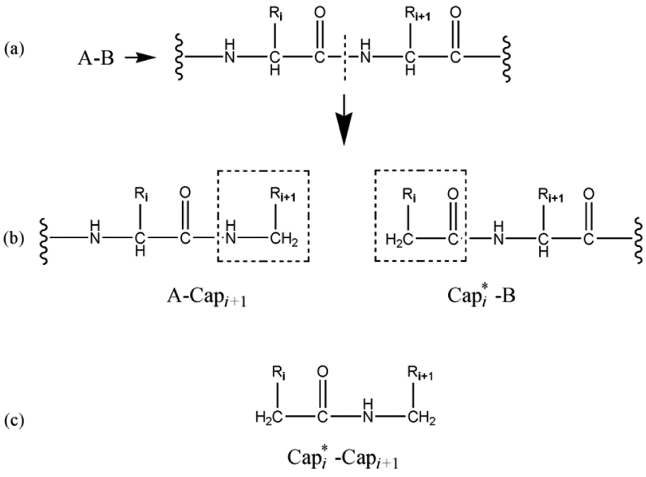


Figure 1. MFCC scheme in which the peptide bond is cut in panel a and the fragments are capped with Cap_{i+1} and its conjugate Cap_i^* in panel b, where *i* represents the index of the *i*th amino acid in the given protein. The atomic structure of the concap is shown in panel c. The concap is defined as the $\mathrm{Cap}_i^*{-}\mathrm{Cap}_{i+1}$ fused molecular species.

By cutting the peptide bond of the protein into amino acid fragments and inserting a pair of concaps, CH_2R_iCO- and $-NHCH_2R_{i+1}$ (*i* denotes the index of the *i*th amino acid), at the cutting location to cap the fragments, the interaction energy for the protein−ligand binding system ($E_{\mathrm{P-L}}$) is given by the following expression[16,49−51]

$$E_{\mathrm{P-L}} = \sum_{k=1}^{N-2} E_{\mathrm{F}_k-\mathrm{L}} - \sum_{k=1}^{N-3} E_{\mathrm{CC}_k-\mathrm{L}} - \sum_{k=1}^{N-2} E_{\mathrm{F}_k} + \sum_{k=1}^{N-3} E_{\mathrm{CC}_k} - E_{\mathrm{L}} \quad (1)$$

where $E_{\mathrm{F}_k-\mathrm{L}}$ and $E_{\mathrm{CC}_k-\mathrm{L}}$ represent the total energy of the *k*th capped fragment and ligand and the total energy of the *k*th concap and ligand, respectively, E_{F_k} and E_{CC_k} are the self-energy of the *k*th capped fragment and *k*th concap, respectively, and E_{L} is the energy of the ligand. For a protein with *N* amino acids, there are $N - 2$ capped fragments and $N - 3$ concaps. In situations where the protein has additional chemical bonds between non-neighboring residues such as disulfide bonds, additional cutting of these bonds is needed as described in ref 48. Figure 2 shows the comparison between the full system M062X/6-311G** and MFCC results for calculating the interaction energy between Efavirenz and a polypeptide (containing 19 residues) extracted from HIV-1 reverse transcriptase (RT). The MFCC method can accurately reproduce the *ab initio* interaction energy between the protein and ligand with a low computational cost. Full quantum mechanical studies of the mutational effect in the binding of Nevirapine and Efavirenz to HIV-1 RT based on the MFCC method were presented in refs 50 and 51, respectively.

2.2. Total Energy of the Protein

Using the MFCC approach, the total electron density of protein with *N* amino acids can be obtained by linear combination of individual densities of capped fragments using the MFCC ansatz[17,52,53]

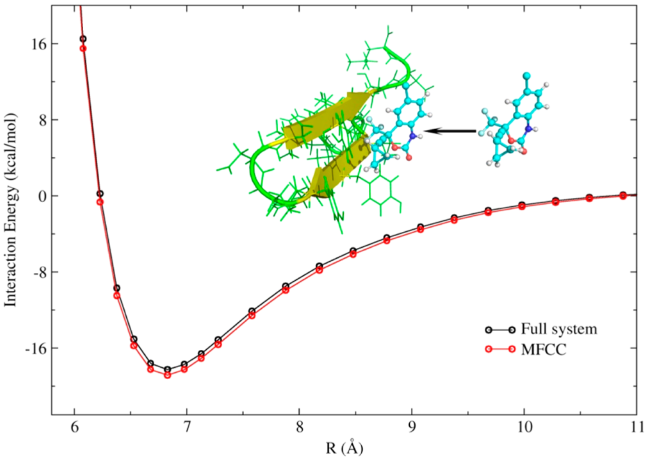


Figure 2. One-dimensional interaction potential curves for Efavirenz and a fraction of HIV-1 reverse transcriptase (Asn175 to Leu193 of chain A) extracted from the RT−Efavirenz complex (Protein Data Bank entry 1FKO). The interaction energy is calculated at the M062X/6-311G** level along the direction from the geometric center of Efavirenz to that of the polypeptide.

$$\rho = \sum_{k=1}^{N-2} \rho_{\mathrm{F}_k} - \sum_{k=1}^{N-3} \rho_{\mathrm{CC}_k} \tag{2}$$

where ρ_{F_k} is the density of the kth protein fragment and ρ_{CC_k} is the density of the kth concap. The same result can be obtained for the electrostatic potential and dipole moment.[17,52−54]

After the density of the full protein system is obtained from the MFCC calculation, one can employ DFT to compute the total energy (E) of the protein using the following equation

$$E(\rho) = T[\rho] - \sum_{\alpha} Z_\alpha \int \frac{\rho(\mathbf{r})}{|\mathbf{R}_\alpha - \mathbf{r}|}\,\mathrm{d}\mathbf{r} + \frac{1}{2}\int \phi(\mathbf{r})\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} + \sum_{\alpha,\beta} \frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} + E_{\mathrm{XC}}[\rho] \tag{3}$$

where $T[\rho]$ is the kinetic energy, $\phi(\mathbf{r})$ is the electrostatic potential (electron contribution only), and $E_{\mathrm{XC}}[\rho]$ is the exchange-correlation energy. The kinetic energy can also be approximately obtained by the MFCC ansatz[17]

$$T[\rho] = \sum_{k=1}^{N-2} T_{\mathrm{F}_k}[\rho_{\mathrm{F}_k}] - \sum_{k=1}^{N-3} T_{\mathrm{CC}_k}[\rho_{\mathrm{CC}_k}] \tag{4}$$

More details about integrating eq 3 can be found in ref 17.

Another approach is to compute the protein energy by constructing the density matrix (DM) of the system based on fragment density matrices. The density matrix of the molecular system can be obtained using the MFCC-DM method[19]

$$P^{\mu\nu} = \sum_{k=1}^{N-2} P^{\mu\nu}_{\mathrm{F}_k} - \sum_{k=1}^{N-3} P^{\mu\nu}_{\mathrm{CC}_k} \tag{5}$$

where $P^{\mu\nu}_{\mathrm{F}_k}$ and $P^{\mu\nu}_{\mathrm{CC}_k}$ are the density matrix elements of the kth fragment and kth concap, respectively. After the density matrix of the full protein system is obtained, the total HF or DFT energy of the protein can be calculated directly from the full density matrix.[19]

A more accurate treatment to include some non-zero off-diagonal density matrix elements to account for the close contact interactions is introduced by pairwise interaction correction (PIC).[20] In the MFCC-DM-PIC approach, two residues that are not simultaneously present in a fragment and are within a certain distance of each other are paired, which is termed an interacting unit. The PICs on the density matrix element associated with the interacting units (residues i and j) are obtained by the following relation

$$P^{\mu\nu}_{\mathrm{PIC}}(i-j) = P^{\mu\nu}_{ij} - P^{\mu\nu}_{i} - P^{\mu\nu}_{j} \tag{6}$$

and they are added to the total density matrix of the full system from the MFCC-DM calculation. The PIC-corrected density matrix is then used to calculate the molecular properties and total energy of the protein.[20]

3. EE-GMFCC METHOD

3.1. Total Energy of the Protein

In the EE-GMFCC scheme, each capped fragment calculation is embedded in the electrostatic field of the point charges representing the remaining amino acids in the protein, which accounts for the electronic polarization effect of the protein environment and is also a key difference from the original MFCC approach. Moreover, generalized concaps (Gconcaps) are introduced to include the two-body QM interaction energies between sequentially non-neighboring fragments that are spatially in close contact (see Figure 3). If the minimal distance between two non-neighboring residues i and j is within a predefined distance threshold λ (normally, λ is set to 4 Å[41]), these two residues are considered to be in close contact (defined as Gconcap) and their interaction is calculated at the QM level. The total energy of the protein (with N amino acids and N_{GC} Gconcaps) using the EE-GMFCC method can be expressed as[41,42]

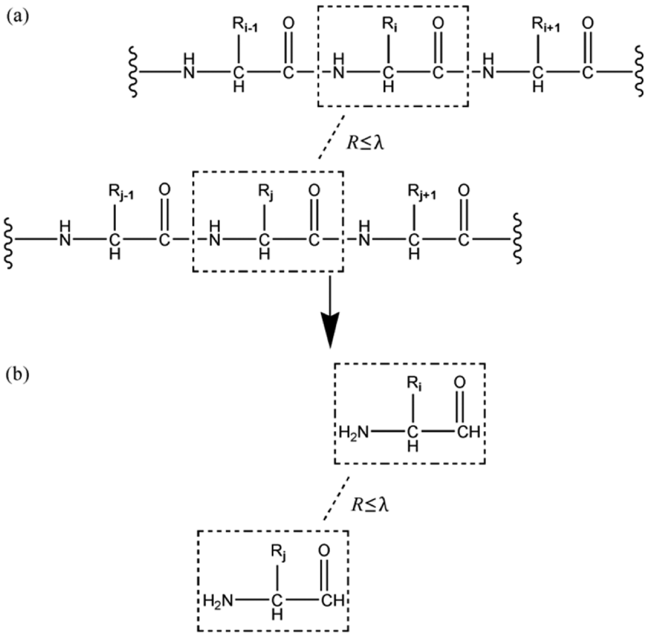


Figure 3. (a) Generalized concap (Gconcap) scheme, when the distance (R) between two non-neighboring residues i and j is within a distance threshold λ ($R \le \lambda$). (b) Atomic structure of Gconcap.

$$E = \sum_{k=1}^{N-2} \tilde{E}_{\mathrm{F}_k} - \sum_{k=1}^{N-3} \tilde{E}_{\mathrm{CC}_k} + \sum_{k=1}^{N_{\mathrm{GC}}} (\tilde{E}_{ij}^{k} - \tilde{E}_{i}^{k} - \tilde{E}_{j}^{k}) - E_{\mathrm{DC}} \tag{7}$$

where $\tilde{E}$ denotes the sum of the self-energy of the fragment and the interaction energy between the fragment and background charges of the remaining system. $\tilde{E}_{ij}^{k} - \tilde{E}_{i}^{k} - \tilde{E}_{j}^{k}$ represents the two-body QM interaction energy between residues i and j in the kth Gconcap. E_{DC} is the interaction energy doubly counted in the previous terms of eq 7, which is approximated by the pairwise charge−charge interactions. The complete definition of E_{DC} can be found in ref 41.

The critical aspect of the EE-GMFCC method is the electrostatic embedding scheme for each fragment calculation, which ensures the electronic polarization effect is properly taken into account. Through the embedding scheme, many-body Coulomb effects from other parts of the protein are included at the HF or DFT level. Typically, the fragment using the EE-GMFCC method consists of fewer than 65 atoms,[41,42] which makes high-level *ab initio* methods applicable for proteins.

Numerical studies were performed to calculate the total energies of 18 globular proteins (containing 243−1142 atoms) using the EE-GMFCC approach at the HF/6-31G* level.[41] The total energies calculated by the EE-GMFCC approach show excellent agreement with the full system results. The overall mean unsigned error (MUE) of EE-GMFCC for the 18 proteins is 2.39 kcal/mol with respect to the full system HF/6-31G* calculations when the distance threshold λ was set to 4.0 Å.[41] In addition, the EE-GMFCC approach was applied to proteins at the DFT and MP2 level, also showing deviations of only a few kilocalories per mole from the corresponding full system results. Figure 4 shows the relative energy profile of 19 conformers for one globular protein [Protein Data Bank (PDB) entry 2KCF, 576 atoms] at the HF/6-31G* level. The conformers were extracted from molecular dynamics (MD) simulation. As shown in Figure 4, the relative EE-GMFCC energies agree well with the full system results. Figure 5 shows a comparison of CPU time between the standard MP2 and EE-GMFCC calculations.

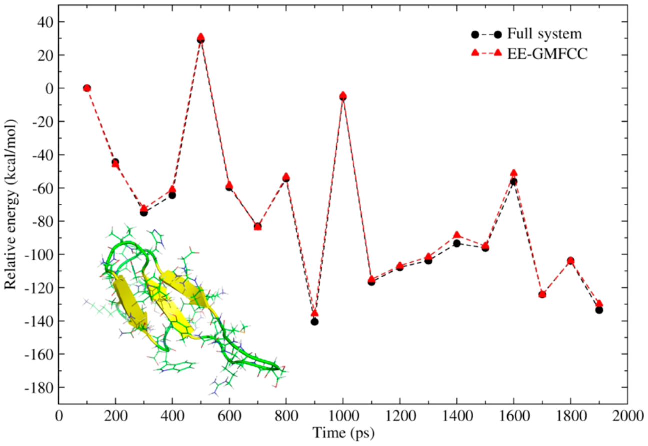


Figure 4. Comparison of the relative energies of 19 conformers (PDB entry 2KCF, 576 atoms) selected from a 2 ns MD simulation between the standard full system HF/6-31G* calculations (black circles) and EE-GMFCC results (red triangles).

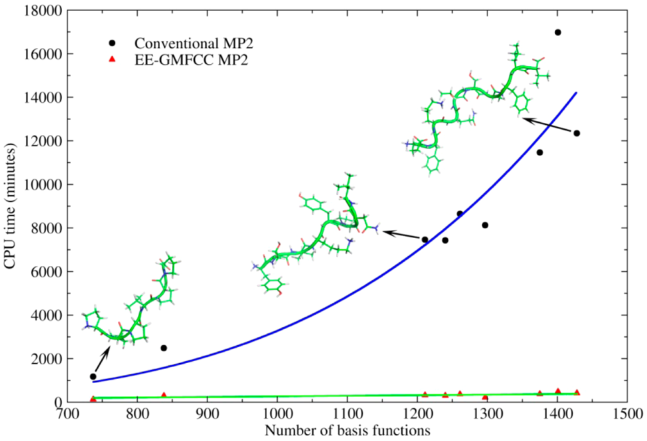


Figure 5. CPU time for the full system and EE-GMFCC calculations as a function of the number of basis functions at the MP2/6-31G* level.

3.2. Electron Density and Electrostatic Potential

On the basis of the EE-GMFCC method, molecular properties (Q) for a given protein with N amino acids, such as the dipole moment (μ), electron density (ρ),[52−54] and electrostatic potential (ϕ),[25,54−57] can also be obtained using the following expression[42]

$$Q = \sum_{k=1}^{N-2} Q_{\mathrm{F}_k} - \sum_{k=1}^{N-3} Q_{\mathrm{CC}_k} + \sum_{k=1}^{N_{\mathrm{GC}}} (Q_{ij}^{k} - Q_{i}^{k} - Q_{j}^{k}) \tag{8}$$

where the notations are similar to those described for eq 7, and the molecular property Q for each fragment is also calculated in the embedding electrostatic field of the remaining system. Figure 6 shows the electrostatic potential and dipole moment of a globular protein (PDB entry 2LAJ, 666 atoms) calculated by EE-GMFCC and full system M062X/6-31G*, respectively. The EE-GMFCC method gives accurate molecular properties for direct linear-scaling computation of protein systems. The electron density and electrostatic potential calculated at the *ab initio* level could improve the accuracy for protein X-ray structure refinement[58] and the prediction of vibrational Stark shifts at the active site of enzymes,[59] respectively.

3.3. Solvation Energy of the Protein

The EE-GMFCC method can be combined with the conductor-like polarizable continuum model (CPCM[60,61]), termed EE-GMFCC−CPCM, for *ab initio* calculation of the electrostatic solvation energies of proteins. The details of the EE-GMFCC−CPCM approach can be found in ref 42. For 12 proteins (containing 218−803 atoms), the electrostatic solvation energies [G(ele)] calculated by EE-GMFCC−CPCM are in good agreement with the full sytem HF/6-31G* calculations.[42] The MUE of G(ele) determined by EE-GMFCC−CPCM is only 3.36 kcal/mol with respect to the full system results for those 12 proteins. The relative G(ele) of 19 different conformations of one small protein (PDB entry 2I9M) are also calculated using EE-GMFCC−CPCM and compared with full system results (see Figure 7). As one can see from Figure 7, the G(ele) of 2I9M undergoes a large fluctuation between −470 and −300 kcal/mol for different conformers. The MUE of the relative G(ele) calculated by EE-GMFCC−CPCM is merely 1.02 kcal/mol with respect to full system energies for 2I9M.

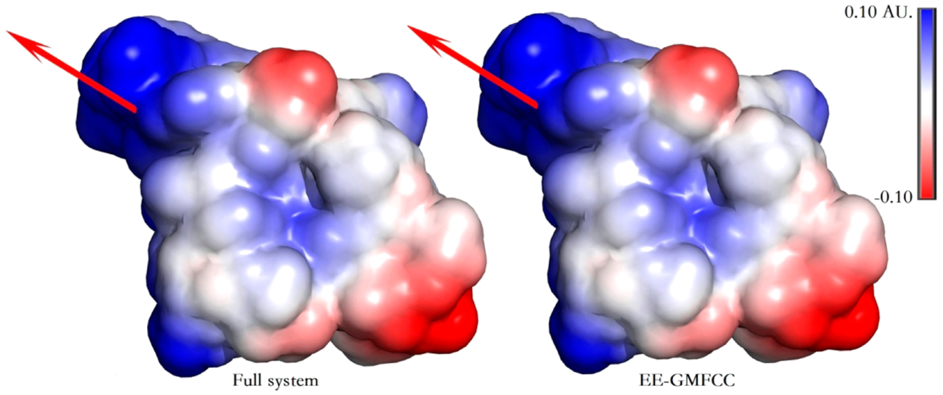


Figure 6. Electrostatic potential in atomic units (AU) at the solvent-accessible surface of the protein (PDB entry 2LAJ, 666 atoms) from full system (left) and EE-GMFCC (right) calculations at the M062X/6-31G* level. The red arrows represent the calculated dipole moments using full system (295.258 D) and EE-GMFCC (297.878 D) methods.

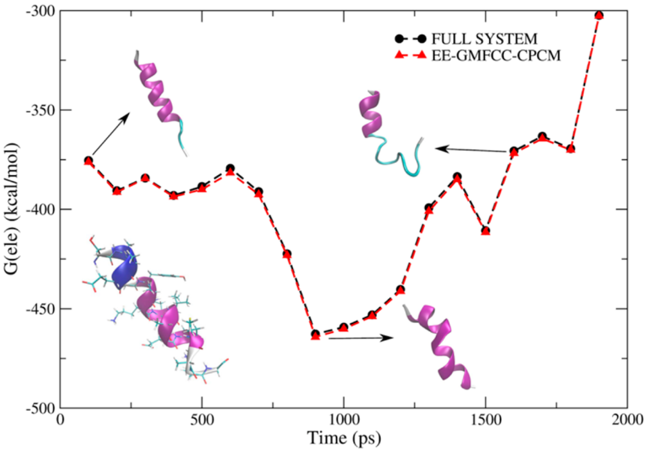


Figure 7. Variation of the electrostatic solvation energy [G(ele)] over 19 different conformations selected from a 2 ns MD simulation for the protein of PDB entry 2I9M. Red triangles and black circles represent the results calculated using the EE-GMFCC−CPCM and standard full system HF/6-31G* methods, respectively.

Figure 8 shows a comparison of CPU time for EE-GMFCC−CPCM and full system HF/6-31G* calculations on 11 small proteins. As expected, the computational time scale for the EE-GMFCC−CPCM calculation is O(N) with a low prefactor, in contrast to O(N^3) for the conventional HF/CPCM calculation of the entire system.

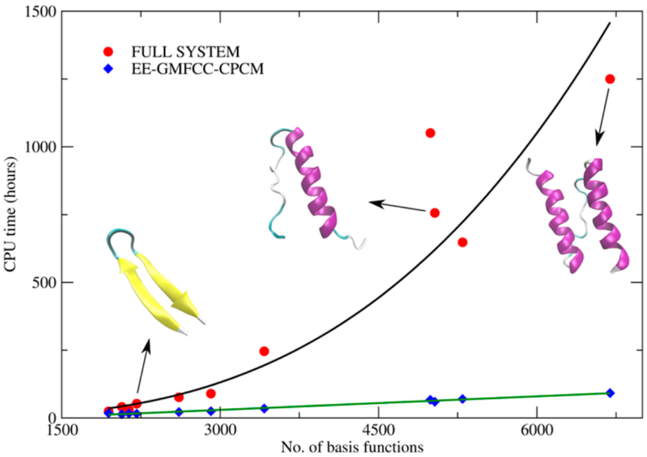


Figure 8. CPU time for the full system and EE-GMFCC−CPCM calculations as a function of the number of basis functions at the HF/6-31G* level.

4. AF-QM/MM METHOD FOR NMR CHEMICAL SHIFTS

4.1. Calculating Protein NMR Chemical Shifts in the Gas Phase

NMR spectroscopy is an invaluable and widely used technique in chemistry and biology. For proteins, the chemical shift tensors are key parameters in the NMR experiment, allowing signals from different nuclei of any given type in a molecule to be distinguished. Although the chemical shifts are the most precise NMR parameters that can be obtained for biomolecules, the inherently complex dependency on geometric, dynamic, and electronic properties has made accurate prediction of protein chemical shifts a significant challenge.[62]

Over the past two decades, QM methods have become increasingly useful tools for NMR chemical shift prediction.[63] However, because of the poor scaling of *ab initio* methods, it has not been practical to apply standard all-electron quantum chemistry methods to large proteins. Because the nuclear shielding is fundamentally a local physical property, several fragmentation methods have been proposed for protein NMR chemical shift calculation at the *ab initio* levels.[64−68] In our previous studies,[38−40] an efficient AF-QM/MM approach was shown to be applicable to routine *ab initio* NMR chemical shift calculations for proteins of any size.

The basic fragmentation scheme of the AF-QM/MM approach is shown in Figure 9a. In this method, the entire protein system is divided into nonoverlapping fragments termed core regions. Usually, we take each amino acid as a core region. The residues within a certain range of the core region are assigned as the buffer region. Both the core region and its buffer region are treated by QM, whereas the rest of the system is described by background charges. The purpose of the buffer area is to include the local QM effects on the chemical shifts of the core region. Each fragment-centric QM/MM calculation is conducted separately. All the fragment-based calculations are mutually independent and parallelizable. At the end, only the shielding constants of the atoms in the core region are extracted from the individual QM/MM calculations. A more detailed illustration of the automated fragmentation scheme is presented in Figure 9b. The detailed distance-

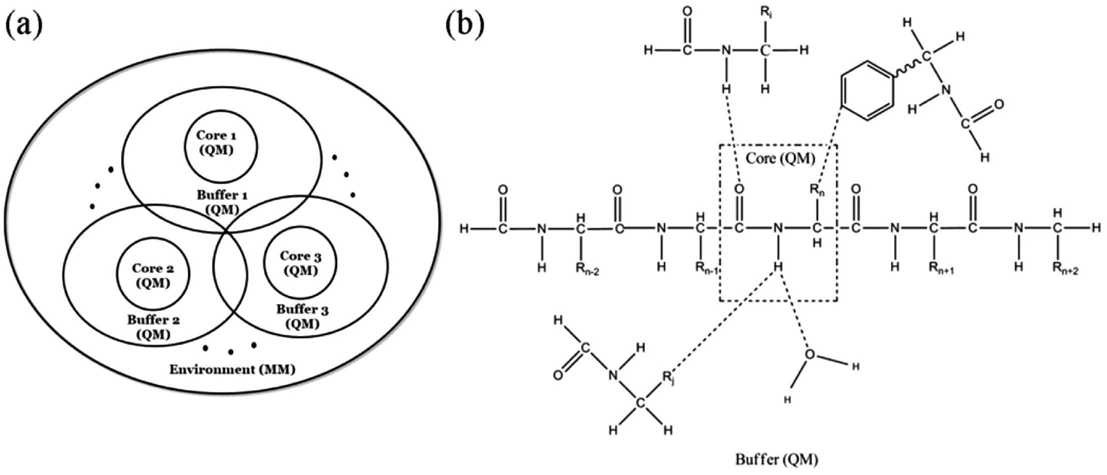


Figure 9. (a) Subsetting scheme for the AF-QM/MM approach. (b) If the *n*th residue is the core region, the sequentially connected (*n* − 2)th, (*n* − 1)th, (*n* + 1)th, and (*n* + 2)th residues are included in the buffer region. In addition, the residues in spatial contact with the *n*th residue are also assigned to the buffer region.

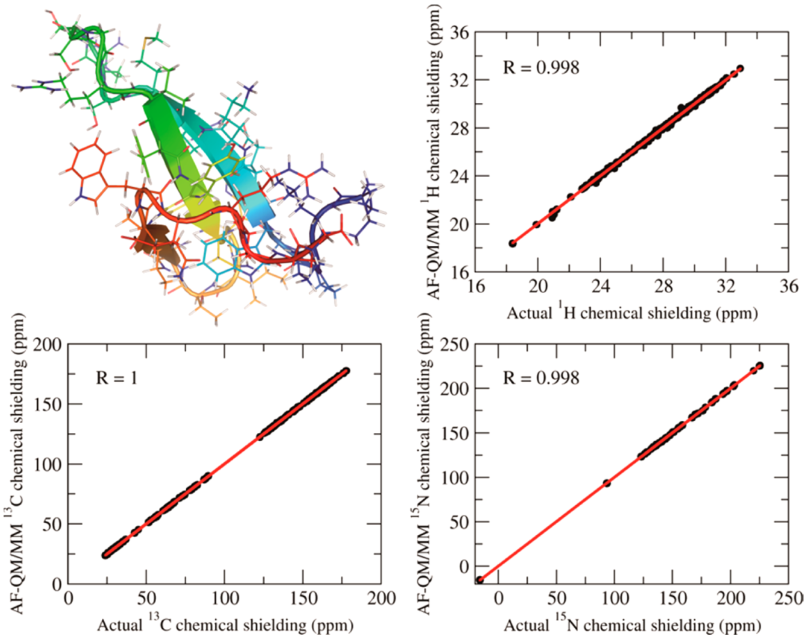


Figure 10. Three-dimensional structure of the Pin1 WW domain (PDB entry 1PIN, 558 atoms) and the correlation between AF-QM/MM and full system B3LYP/6-31G** calculations (in the gas phase) for ^{1}H, ^{13}C, and ^{15}N chemical shielding.

dependent criteria for assigning the buffer region are described in refs 38 and 39.

The results of the AF-QM/MM method shows excellent agreement with the results of the full system (gas phase) calculations of NMR chemical shieldings for the Pin1 WW domain. As shown in Figure 10, the correlation coefficients between AF-QM/MM and full system B3LYP/6-31G** calculations are almost 1 for ^{1}H, ^{13}C, and ^{15}N chemical shieldings. The root-mean-square errors (rmses) for ^{1}H, ^{13}C, and ^{15}N chemical shieldings are 0.09, 0.23, and 0.40 ppm, respectively, with respect to full system results. The environmental electrostatic potential is also indispensable for accurately reproducing the chemical shielding using the fragmentation approach.[38] The AF-QM/MM method has also been successfully applied to predicting the chemical shift anisotropies in proteins[44] and vicinal *J* spin−spin coupling constants for the protein backbone.[43]

4.2. Calculating Protein NMR Chemical Shifts in Implicit Solvent

As most NMR measurements are performed on liquid samples and NMR chemical shifts are quite sensitive to the solvent effects, the AF-QM/MM approach can be improved by incorporating the implicit solvent model to calculate protein NMR chemical shifts in solution.[39] In the continuum-solvent model, the solute (protein) is represented by a charge distribution $\rho(\mathbf{r})$ embedded in a cavity surrounded by a polarizable medium with dielectric constant ε. The solute charge distribution polarizes the dielectric medium and creates a reaction field that acts back to polarize the solute until equilibrium is reached. The reaction field acting on the solute can be effectively represented by that of induced charges on the cavity surface according to the classical electrostatic theory. On top of the AF-QM/MM method, we use the DivCon program[69] that combines the linear-scaling divide-and-conquer semi-

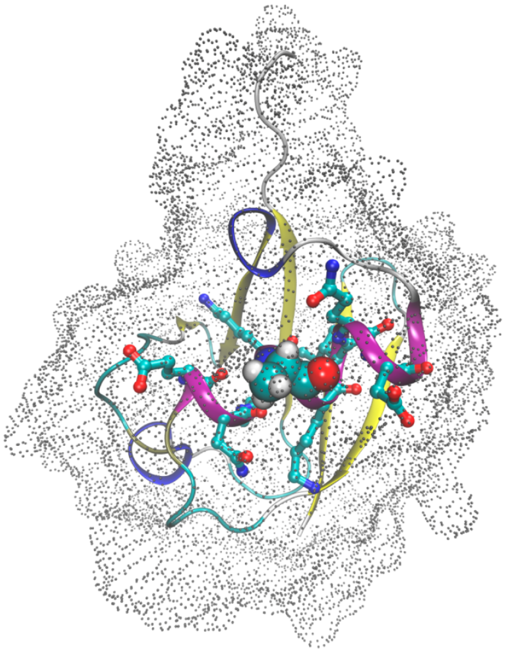


Figure 11. X-ray structure of ubiquitin (PDB entry 1UBQ, 1231 atoms) together with the surface charges calculated by DivCon. The core QM region and buffer region are represented by a ball-and-stick model and a stick model, respectively. The rest of the protein is treated with the point charge model.

empirical algorithm with the Poisson−Boltzmann (PB) equation to perform the self-consistent reaction field (SCRF) calculation. Then the set of point charges of the MM environment and on the molecular surface (derived from SCRF calculations) that represents the reaction field are used as the background charges in the QM calculation as shown in Figure 11.

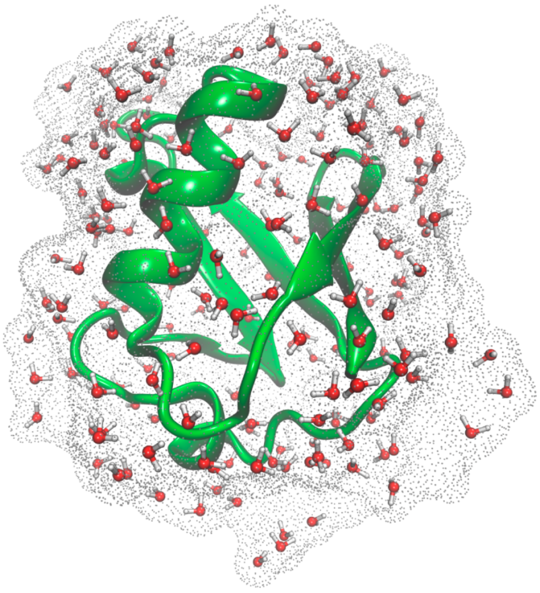


Figure 13. Graphic representation of ubiquitin together with the explicit water molecules and surface charges calculated with DivCon.

AF-QM/MM calculation of NMR chemical shifts for four proteins (Trp-cage, Pin1 WW domain, GB3, and ubiquitin) has been conducted using the B3LYP method.[39] The calculated chemical shifts of ^1H and ^{13}C for these four proteins are in excellent agreement with experimentally measured values and represent a clear improvement over those from the gas phase calculation (see Figure 12), while the nonpolar ^{13}C chemical shifts are less affected by the solvent. However, although the inclusion of the solvent effect also improves the accuracy for ^{15}N chemical shifts, the computed results do not agree with experimental values as well as ^1H and ^{13}C. Furthermore, the

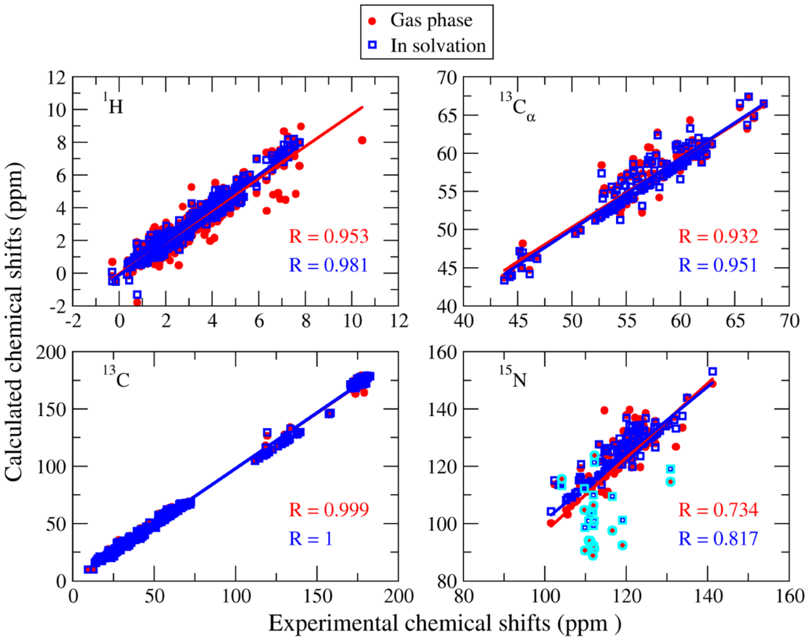


Figure 12. Correlation between experimental ^1H, ^{13}C, $^{13}C_\alpha$, and ^{15}N NMR chemical shifts and calculated chemical shifts of four proteins (Trp-cage, Pin1 WW domain, GB3, and ubiquitin) using the AF-QM/MM method. The results for ^{15}N in the side chains are highlighted with cyan circles. The amide protons were excluded. The chemical shifts of carbonyl carbons were calculated using the B3LYP functional with the mixed (6-311++G**/4-31G*) basis set, while the chemical shifts for other atoms were computed at the B3LYP/6-31G** level.

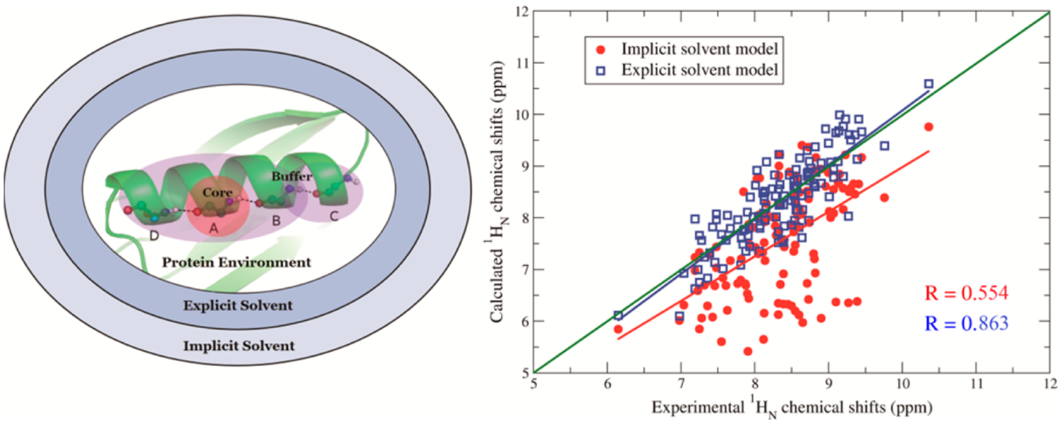


Figure 14. Fragmentation scheme of the AF-QM/MM approach with the explicit solvent model (left). The cooperative hydrogen bonding effect is also taken into account. Correlation between experimental and calculated ^1H_N chemical shifts of two proteins (GB3 and ubiquitin) by AF-QM/MM using the B3LYP functional with the mixed (6-311++G**/4−31G*) basis set (right).

AF-QM/MM calculated result can accurately reflect the dependence of $^{13}C_\alpha$ chemical shifts on a protein's secondary structure, and the *ab initio* chemical shifts can be utilized to discriminate the native structure of proteins from decoy conformations through direct comparison between experiment and theory.[39]

4.3. Calculating Protein NMR Chemical Shifts in Explicit Solvent

In AF-QM/MM calculations with the implicit solvent model, the predicted amide proton (^1H_N) chemical shifts of proteins still have large deviations from the experimental values. The specific solvent−solute interactions such as hydrogen bonding between the polar amide group and water molecules cannot be accurately described using implicit solvent. 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.[40] To place the explicit water molecules around the protein, we utilized the PLACEVENT program that is based on the three-dimensional reference interaction site model (3D-RISM).[70] Only the water molecules in the first and second solvation shells (within 6.0 Å of any atom of the protein) are explicitly treated as part of the entire system (see Figure 13), while the implicit solvent model was used to represent the bulk solvent effect beyond the second solvent shell. As shown in Figure 14, the calculated ^1H_N chemical shifts of two proteins (GB3 and ubiquitin) using the explicit solvent model show remarkable improvement over those from the implicit solvation calculation.[40]

5. CONCLUDING REMARKS

The EE-GMFCC approach is an accurate and efficient method for QM calculation of the molecular properties, total energy, and electrostatic solvation energy of proteins. The molecular properties calculated at the *ab initio* level could greatly improve the accuracy of protein X-ray structure refinement and predict the vibrational Stark shift at the active site of enzymes. The EE-GMFCC method is linear-scaling with a small prefactor, trivially parallel, and can be readily applied in performing protein−ligand binding affinity calculations in solution, structural optimization of proteins, and molecular dynamics simulation with high-level *ab initio* electronic structure theories.

Via combination of implicit and explicit solvent models, the protein NMR chemical shifts calculated by the linear-scaling AF-QM/MM method are in excellent agreement with experimental values. Other NMR parameters, such as the chemical shift anisotropy tensor and spin−spin coupling constant, have also been studied by using this approach.[43,44] The applications of the AF-QM/MM method may also be extended to more general biological systems,[45] such as DNA/RNA, metalloprotein, protein−ligand, and membrane protein−lipid complexes.

AUTHOR INFORMATION

Corresponding Authors

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

Notes

The authors declare no competing financial interest.

Biographies

Xiao He earned a B.S. in physics (2003) and a M.S. in chemistry (2006) from Nanjing University and a Ph.D. in chemistry (2010) from the University of Florida under the supervision of Professor Kenneth Merz. He was trained as a postdoctoral researcher at the University of Illinois at Urbana-Champaign (2011−2012), where his advisor was Professor So Hirata. He is currently an Associate Professor at the State Key Laboratory of Precision Spectroscopy at East China Normal University.

Tong Zhu earned a B.S. in physics (2007) and a M.S. in physics (2010) from Shandong Normal University and a Ph.D. in physics (2013) from East China Normal University under the supervision of Professor Xiao He and John Z. H. Zhang. He is currently a postdoctoral researcher at East China Normal University.

Xianwei Wang earned a B.S. in physics (2009) from Shandong Normal University. He is currently a fifth-year graduate student at East China Normal University.

Jinfeng Liu earned a B.S. in physics (2011) from Shandong Normal University. He is currently a third-year graduate student at East China Normal University.

John Z. H. Zhang earned a B.S. in physics (1982) from East China Normal University and a Ph.D. in chemistry (1987) from the University of Houston. He was trained as a postdoctoral researcher at the University of California at Berkeley (1987−1990). He then joined the chemistry faculty of New York University, where he became full professor in 1997. He is currently a professor of East China Normal University and New York University Shanghai.

■ ACKNOWLEDGMENTS

We thank the National Natural Science Foundation of China (Grants 10974054, 20933002, and 21303057) and the Shanghai PuJiang program (09PJ1404000). X.H. is also supported by the Specialized Research Fund for the Doctoral Program of Higher Education (Grant 20130076120019) and the Fundamental Research Funds for the Central Universities. We thank the Supercomputer Center of East China Normal University for providing us computational time. Acknowledgment is extended to Xiangyu Jia and Prof. Ye Mei for their contributions in the EE-GMFCC−CPCM work.

■ REFERENCES

(1) Szabo, A.; Ostlund, N. S. *Modern quantum chemistry: Introduction to advanced electronic structure theory*, 1st ed.; McGraw-Hill: New York, 1989.

(2) Goedecker, S. Linear scaling electronic structure methods. *Rev. Mod. Phys.* **1999**, *71*, 1085−1123.

(3) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Headgordon, M. The Continuous Fast Multipole Method. *Chem. Phys. Lett.* **1994**, *230*, 8−16.

(4) Friesner, R. A.; Murphy, R. B.; Beachy, M. D.; Ringnalda, M. N.; Pollard, W. T.; Dunietz, B. D.; Cao, Y. X. Correlated ab initio electronic structure calculations for large molecules. *J. Phys. Chem. A* **1999**, *103*, 1913−1928.

(5) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. Achieving linear scaling for the electronic quantum coulomb problem. *Science* **1996**, *271*, 51−53.

(6) Yang, W. T. Direct Calculation of Electron-Density in Density-Functional Theory. *Phys. Rev. Lett.* **1991**, *66*, 1438−1441.

(7) Hampel, C.; Werner, H. J. Local treatment of electron correlation in coupled cluster theory. *J. Chem. Phys.* **1996**, *104*, 6286−6297.

(8) Saebø, S.; Pulay, P. Local Treatment of Electron Correlation. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213−236.

(9) He, X.; Merz, K. M., Jr. Divide and Conquer Hartree-Fock Calculations on Proteins. *J. Chem. Theory Comput.* **2010**, *6*, 405−411.

(10) He, X.; Sode, O.; Xantheas, S. S.; Hirata, S. Second-order many-body perturbation study of ice Ih. *J. Chem. Phys.* **2012**, *137*, 204505.

(11) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112*, 632−672.

(12) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: An approximate computational method for large molecules. *Chem. Phys. Lett.* **1999**, *313*, 701−706.

(13) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment molecular orbital method: Use of approximate electrostatic potential. *Chem. Phys. Lett.* **2002**, *351*, 475−480.

(14) Fedorov, D. G.; Kitaura, K. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A* **2007**, *111*, 6904−6914.

(15) He, X.; Fusti-Molnar, L.; Cui, G. L.; Merz, K. M. Importance of Dispersion and Electron Correlation in ab Initio Protein Folding. *J. Phys. Chem. B* **2009**, *113*, 5290−5300.

(16) Zhang, D. W.; Zhang, J. Z. H. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein-molecule interaction energy. *J. Chem. Phys.* **2003**, *119*, 3599−3605.

(17) He, X.; Zhang, J. Z. H. A new method for direct calculation of total energy of protein. *J. Chem. Phys.* **2005**, *122*, 031103.

(18) He, X.; Zhang, J. Z. H. The generalized molecular fractionation with conjugate caps/molecular mechanics method for direct calculation of protein energy. *J. Chem. Phys.* **2006**, *124*, 184703.

(19) Chen, X. H.; Zhang, Y. K.; Zhang, J. Z. H. An efficient approach for ab initio energy calculation of biopolymers. *J. Chem. Phys.* **2005**, *122*, 184105.

(20) Chen, X. H.; Zhang, J. Z. H. Molecular fractionation with conjugated caps density matrix with pairwise interaction correction for protein energy calculation. *J. Chem. Phys.* **2006**, *125*, 044903.

(21) Mei, Y.; He, X.; Ji, C. G.; Zhang, D. W.; Zhang, J. Z. H. A Fragmentation Approach to Quantum Calculation of Large Molecular Systems. *Prog. Chem.* **2012**, *24*, 1058−1064.

(22) Deev, V.; Collins, M. A. Approximate ab initio energies by systematic molecular fragmentation. *J. Chem. Phys.* **2005**, *122*, 154102.

(23) Collins, M. A.; Deev, V. A. Accuracy and efficiency of electronic energies from systematic molecular fragmentation. *J. Chem. Phys.* **2006**, *125*, 104104.

(24) 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*, 10040−10049.

(25) Exner, T. E.; Mezey, P. G. Ab initio-quality electrostatic potentials for proteins: An application of the ADMA approach. *J. Phys. Chem. A* **2002**, *106*, 11791−11800.

(26) Exner, T. E.; Mezey, P. G. Ab initio quality properties for macromolecules using the ADMA approach. *J. Comput. Chem.* **2003**, *24*, 1980−1986.

(27) Exner, T. E.; Mezey, P. G. The field-adapted ADMA approach: Introducing point charges. *J. Phys. Chem. A* **2004**, *108*, 4301−4309.

(28) Babu, K.; Gadre, S. R. Ab initio quality one-electron properties of large molecules: Development and testing of molecular tailoring approach. *J. Comput. Chem.* **2003**, *24*, 484−495.

(29) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. Molecular tailoring approach for geometry optimization of large molecules: Energy evaluation and parallelization strategies. *J. Chem. Phys.* **2006**, *125*, 104109.

(30) Isegawa, M.; Wang, B.; Truhlar, D. G. Electrostatically Embedded Molecular Tailoring Approach and Validation for Peptides. *J. Chem. Theory Comput.* **2013**, *9*, 1381−1393.

(31) Li, S. H.; Li, W.; Fang, T. An efficient fragment-based approach for predicting the ground-state energies and structures of large molecules. *J. Am. Chem. Soc.* **2005**, *127*, 7215−7226.

(32) Li, W.; Li, S. H.; Jiang, Y. S. Generalized energy-based fragmentation approach for computing the ground-state energies and properties of large molecules. *J. Phys. Chem. A* **2007**, *111*, 2193−2199.

(33) Richard, R. M.; Herbert, J. M. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012**, *137*, 064113.

(34) Dahlke, E. E.; Truhlar, D. G. Electrostatically embedded many-body expansion for large systems, with applications to water clusters. *J. Chem. Theory Comput.* **2007**, *3*, 46−53.

(35) Dahlke, E. E.; Truhlar, D. G. Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Møller-Plesset perturbation theory energies for large water clusters. *J. Chem. Theory Comput.* **2007**, *3*, 1342−1348.

(36) Xie, W. S.; Gao, J. L. Design of a next generation force field: The X-POL potential. *J. Chem. Theory Comput.* **2007**, *3*, 1890−1900.

(37) Xie, W. S.; Song, L. C.; Truhlar, D. G.; Gao, J. L. The variational explicit polarization potential and analytical first derivative of energy: Towards a next generation force field. *J. Chem. Phys.* **2008**, *128*, 234108.

(38) 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*, 10380−10388.

(39) Zhu, T.; He, X.; Zhang, J. Z. H. Fragment density functional theory calculation of NMR chemical shifts for proteins with implicit solvation. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7837−7845.

(40) 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.

(41) Wang, X. W.; Liu, J. F.; Zhang, J. Z. H.; He, X. Electrostatically Embedded Generalized Molecular Fractionation with Conjugate Caps

Method for Full Quantum Mechanical Calculation of Protein Energy. *J. Phys. Chem. A* **2013**, *117*, 7149−7161.

(42) Jia, X. Y.; Wang, X. W.; Liu, J. F.; Zhang, J. Z. H.; Mei, Y.; He, X. An improved fragment-based quantum mechanical method for calculation of electrostatic solvation energy of proteins. *J. Chem. Phys.* **2013**, *139*, 214104.

(43) Wang, B.; He, X.; Merz, K. M. Quantum Mechanical Study of Vicinal J Spin-Spin Coupling Constants for the Protein Backbone. *J. Chem. Theory Comput.* **2013**, *9*, 4653−4659.

(44) Tang, S. S.; Case, D. A. Calculation of chemical shift anisotropy in proteins. *J. Biomol. NMR* **2011**, *51*, 303−312.

(45) Case, D. A. Chemical shifts in biomolecules. *Curr. Opin. Struct. Biol.* **2013**, *23*, 172−176.

(46) Zhang, D. W.; Zhang, J. Z. H. Full ab initio computation of protein-water interaction energies. *J. Theor. Comput. Chem.* **2004**, *3*, 43−49.

(47) Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. Fully quantum mechanical energy optimization for protein-ligand structure. *J. Comput. Chem.* **2004**, *25*, 1431−1437.

(48) Chen, X. H.; Zhang, D. W.; Zhang, J. Z. H. Fractionation of peptide with disulfide bond for quantum mechanical calculation of interaction energy with molecules. *J. Chem. Phys.* **2004**, *120*, 839−844.

(49) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. Quantum mechanical map for protein-ligand binding with application to *β*-trypsin/benzamidine complex. *J. Chem. Phys.* **2004**, *120*, 1145−1148.

(50) He, X.; Mei, Y.; Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. Quantum computational analysis for drug resistance of HIV-1 reverse transcriptase to nevirapine through point mutations. *Proteins* **2005**, *61*, 423−432.

(51) Mei, Y.; He, X.; Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. Quantum study of mutational effect in binding of efavirenz to HIV-1 RT. *Proteins* **2005**, *59*, 489−495.

(52) Gao, A. M.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. K. An efficient linear scaling method for ab initio calculation of electron density of proteins. *Chem. Phys. Lett.* **2004**, *394*, 293−297.

(53) Mei, Y.; Zhang, D. W.; Zhang, J. Z. H. New method for direct linear-scaling calculation of electron density of proteins. *J. Phys. Chem. A* **2005**, *109*, 2−5.

(54) Mei, Y.; Wu, E. L.; Han, K. L.; Zhang, J. Z. H. Treating hydrogen bonding in ab initio calculation of biopolymers. *Int. J. Quantum Chem.* **2006**, *106*, 1267−1276.

(55) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. Molecular Tailoring Approach for Simulation of Electrostatic Properties. *J. Phys. Chem.* **1994**, *98*, 9165−9169.

(56) Le, H. A.; Lee, A. M.; Bettens, R. P. A. Accurately Reproducing Ab Initio Electrostatic Potentials with Multipoles and Fragmentation. *J. Phys. Chem. A* **2009**, *113*, 10527−10533.

(57) Reid, D. M.; Collins, M. A. Molecular electrostatic potentials by systematic molecular fragmentation. *J. Chem. Phys.* **2013**, *139*, 184117.

(58) Li, X.; He, X.; Wang, B.; Merz, K. Conformational Variability of Benzamidinium-Based Inhibitors. *J. Am. Chem. Soc.* **2009**, *131*, 7742−7754.

(59) Wang, X. W.; He, X.; Zhang, J. Z. H. Predicting Mutation-Induced Stark Shifts in the Active Site of a Protein with a Polarized Force Field. *J. Phys. Chem. A* **2013**, *117*, 6015−6023.

(60) Klamt, A.; Schüürmann, G. Cosmo: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799−805.

(61) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **2005**, *105*, 2999−3093.

(62) Helgaker, T.; Jaszunski, M.; Ruud, K. Ab initio methods for the calculation of NMR shielding and indirect spin-spin coupling constants. *Chem. Rev.* **1999**, *99*, 293−352.

(63) 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*, 1491−1496.

(64) Frank, A.; Onila, I.; Möller, H. M.; Exner, T. E. Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins. *Proteins* **2011**, *79*, 2189−2202.

(65) Frank, A.; Möller, H. M.; Exner, T. E. Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 2. Level of Theory, Basis Set, and Solvents Model Dependence. *J. Chem. Theory Comput.* **2012**, *8*, 1480−1492.

(66) 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*, 331−339.

(67) 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*, 1428−1444.

(68) Flaig, D.; Beer, M.; Ochsenfeld, C. Convergence of Electronic Structure with the Size of the QM Region: Example of QM/MM NMR Shieldings. *J. Chem. Theory Comput.* **2012**, *8*, 2260−2271.

(69) 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.

(70) Sindhikara, D. J.; Yoshida, N.; Hirata, F. Placevent: An algorithm for prediction of explicit solvent atom distribution: Application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem.* **2012**, *33*, 1536−1543.