AIP | The Journal of Chemical Physics

# An improved fragment-based quantum mechanical method for calculation of electrostatic solvation energy of proteins

Xiangyu Jia, Xianwei Wang, Jinfeng Liu, John Z. H. Zhang, Ye Mei, and Xiao He

Citation: The Journal of Chemical Physics **139**, 214104 (2013); doi: 10.1063/1.4833678

View online: http://dx.doi.org/10.1063/1.4833678

View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/139/21?ver=pdfcov

Published by the AIP Publishing

AIP Publishing Re-register for Table of Content Alerts

Create a profile. Sign up today!

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

# An improved fragment-based quantum mechanical method for calculation of electrostatic solvation energy of proteins

Xiangyu Jia,[1] Xianwei Wang,[1] Jinfeng Liu,[1] John Z. H. Zhang,[1,2] Ye Mei,[1,a)] and Xiao He[1,a)]

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

[2] *Joint Research Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China*

(Received 2 August 2013; accepted 12 November 2013; published online 3 December 2013)

An efficient approach that combines the electrostatically embedded generalized molecular fractionation with conjugate caps (EE-GMFCC) method with conductor-like polarizable continuum model (CPCM), termed EE-GMFCC-CPCM, is developed for *ab initio* calculation of the electrostatic solvation energy of proteins. Compared with the previous MFCC-CPCM study [Y. Mei, C. G. Ji, and J. Z. H. Zhang, J. Chem. Phys. **125**, 094906 (2006)], quantum mechanical (QM) calculation is applied to deal with short-range non-neighboring interactions replacing the classical treatment. Numerical studies are carried out for proteins up to 3837 atoms at the HF/6-31G* level. As compared to standard full system CPCM calculations, EE-GMFCC-CPCM shows clear improvement over the MFCC-CPCM method for both the total electrostatic solvation energy and its components (the polarized solute-solvent reaction field energy and wavefunction distortion energy of the solute). For large proteins with 1000–4000 atoms, where the standard full system *ab initio* CPCM calculations are not affordable, the EE-GMFCC-CPCM gives larger relative wavefunction distortion energies and weaker relative electrostatic solvation energies for proteins, as compared to the corresponding energies calculated by the Divide-and-Conquer Poisson-Boltzmann (D&C-PB) method. Notwithstanding, a high correlation between EE-GMFCC-CPCM and D&C-PB is observed. This study demonstrates that the linear-scaling EE-GMFCC-CPCM approach is an accurate and also efficient method for the calculation of electrostatic solvation energy of proteins. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4833678]

## I. INTRODUCTION

For biological macromolecules like proteins and DNAs, their structures and functions are subtly modulated by solute-solvent interaction.[1–4] Much effort has been devoted to the development of methods for the calculation of molecular properties in solution, with particular interest in solvation energy. Rigorous representation of solvation as in the explicit model can provide a microscopic view of the organization and dynamics of the solvent molecules around the solute, which is indispensable for the study of solvent's response to the variation of solute molecules on multiple time scales. However, computation of thermodynamic properties in explicit solvent requires large scale sampling in phase space, which might be too expensive. Another alternative scheme is to treat the solvent as a continuum dielectric medium, in which the solute is embedded.[5–9] A battery of solvent models has been developed, including the density-based polarized continuum model (PCM),[10–15] single- or multicenter multipolar expansion model,[16,17] Generalized Born (GB) approximation,[18–20] and Poisson-Boltzmann (PB) model.[21–23] In the widely used PCM model, induced charges are generated on solute-solvent surface due to the abrupt change of dielectric constant at this surface and these induced charges polarize the electronic structure of the solute in return. This mutual polarization effect is significant for polar macromolecules, such as the protein and DNA, which can only be accurately described by quantum mechanical (QM) calculations.

Although computer technology has made impressive progress in the past few decades, QM calculation for large molecules still presents a grand challenge. Over the past decade, a range of fragment-based QM method for large systems have been proposed, including the fragment molecular orbital (FMO) method,[24–26] molecular tailing approach (MTA),[27–30] systematic fragmentation method (SFM),[31–34] adjustable density matrix assembler (ADMA)[35–37] approach, electrostatically embedded many-body (EE-MB)[38–40] expansion approach, the generalized energy-based fragmentation (GEBF)[41–43] method, and the molecular fractionation with conjugate caps (MFCC) method.[44–50] Detailed introduction of the fragmentation methods can be found in a recent review.[51] The fragmentation methods have been utilized in a bunch of studies of various systems such as water clusters,[52–54] proteins,[55–58] and protein-ligand complexes.[59–64]

Incorporating the fragmentation methods with solvation models made an important pace toward rigorous study of protein in realistic environment.[65–67] Wavefunction distortion energy, which is nonnegligible but missing in classical treatment, can be correctly captured by QM calculations. In

[a)] Authors to whom correspondence should be addressed. Electronic addresses: ymei@phy.ecnu.edu.cn and xiaohe@phy.ecnu.edu.cn

© 2013 AIP Publishing LLC

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

2006, Mei *et al.*[65] proposed a practical MFCC-CPCM method which combined the MFCC approach with the conductor-like PCM model[68,69] to calculate the electrostatic solvation energies of proteins. Comparison between the MFCC-CPCM approach and the standard full system CPCM calculation indicated that there was still room for improvement, especially in the treatment of short-range interactions such as hydrogen bonds. Fedorov *et al.*[66] incorporated PCM into the FMO method and applied it to some polypeptides. It was found that the FMO/PCM errors with respect to the full system calculation were significantly reduced, when the two-body expansion of the electron density was included in describing the PCM potential with a larger size of the monomer.

Recently, an electrostatically embedded generalized molecular fractionation with conjugate caps method, named EE-GMFCC,[70] was developed for efficient linear-scaling QM calculation of protein energy. In the EE-GMFCC scheme, the total energy of protein is calculated by taking a linear combination of the QM energies of the neighboring residues and the two-body QM interaction energy between non-neighboring residues that are spatially in close contact. All the fragment calculations are embedded in a field of point charges representing the remaining protein environment. Excellent agreement between the EE-GMFCC result and those from the standard full system QM energy calculation was obtained. The EE-GMFCC method is different from the GEBF approach,[42] which is another derivative of the MFCC method.[41,51] In the GEBF approach, all the nearby residues with a pre-defined distance from the central residue are regarded as "overlapping caps," which makes each subsystem include not only the bonded residues with the central residue but also the non-bonded residues which are spatially in close contact with the central residue. As demonstrated in the GEBF paper,[42] some subsystems could contain hundreds of atoms for proteins. In contrast, the philosophy of EE-GMFCC is that it only takes the bonded residues with the central residue as the "overlapping caps," while the non-bonded residues within a short distance are treated as two-body QM interaction corrections and the three- and all higher-order Coulomb effects are approximated through the embedding field. Therefore, the size of each fragment in EE-GMFCC is much smaller than that in the GEBF method, resulting in a smaller prefactor for the linear-scaling fragmentation method. The largest size of the fragment using the EE-GMFCC scheme is normally less than 65 atoms,[70] which makes high-level *ab initio* methods applicable for proteins. Another difference between the EE-GMFCC and GEBF method is the charge model used for the embedding field. GEBF fits the atomic charges in a self-consistent fashion, while EE-GMFCC utilizes the fixed AMBER94 charge model to approximate the electrostatic field of the environment, which has been shown to be an effective and also efficient embedding scheme.[70] Moreover, the analytical atomic gradients of EE-GMFCC can be readily obtained based on the fixed charge embedding approach. In this study, we combine the EE-GMFCC method with the CPCM model, denoted as EE-GMFCC-CPCM, to calculate the electrostatic solvation energy of polypeptides and proteins with up to 3837 atoms. In addition, variations of the electrostatic solvation energy over different conformations of one protein are compared between the EE-GMFCC-CPCM, MFCC-CPCM method, and full system calculations.

This paper is organized as follows. Section II provides a detailed description of the EE-GMFCC-CPCM method and specific procedures for numerical calculations. In Sec. III, numerical test calculations of the EE-GMFCC-CPCM approach are carried out on several polypeptides and proteins for comparison with the MFCC-CPCM and corresponding standard full-system CPCM results. Finally, a brief summary is given in Sec. IV.

## II. THEORY AND METHODOLOGY

### A. EE-GMFCC method

Detailed description of the EE-GMFCC method can be found in a recent paper.[70] Here, we just give a brief review. In the framework of EE-GMFCC method, the protein with $N$ residues is divided into $N - 2$ individual fragments by cutting through the peptide bond (excluding the first and the last peptide bonds) and a pair of conjugate caps is used to saturate each fragment. Hydrogen atoms are added to terminate the molecular caps to avoid dangling bonds (see Figure 1(a)). Considering the importance of the two-body interaction,[71,72] rigorous treatment of the short-range non-neighboring interaction is necessary. In the EE-GMFCC scheme, if the minimal distance between non-neighboring fragments $i$ and $j$ falls within a defined distance threshold $\lambda$, these two residues are considered to be in close contact and their interaction is calculated at the QM level (see Figure 1(b)). Each fragment calculation is embedded in the electrostatic field of the point charges representing the remaining fragments in the protein. The total energy of protein using the EE-GMFCC method can be expressed as

$$
\begin{aligned}
\mathrm{E}_{\text{EE-GMFCC}} &= \sum_{i=2}^{N-1} \tilde{E}(\mathrm{Cap}^*_{i-1}\mathrm{A}_i\mathrm{Cap}_{i+1}) \\
&- \sum_{i=2}^{N-2} \tilde{E}(\mathrm{Cap}^*_{i}\mathrm{Cap}_{i+1}) + \sum_{\substack{i,j>i+2\\ |R_i-R_j|\le\lambda}} (\tilde{E}_{ij} - \tilde{E}_i - \tilde{E}_j)_{\mathrm{QM}} \\
&- \left\{ \sum_{k,l}\sum_{m,n} \frac{q_{m(k)}q_{n(l)}}{R_{m(k)n(l)}} - \sum_{\substack{i,j>i+2\\ |R_i-R_j|\le\lambda}} \sum_{m',n'} \frac{q_{m'(i)}q_{n'(j)}}{R_{m'(i)n'(j)}} \right\},
\end{aligned} \tag{1}
$$

where $\tilde{E}$ represents the summation of the self-energy of the fragment and the interaction energy between the fragment and background charges of the remaining system. $i$ denotes the index of the $i$th residue. $\mathrm{Cap}^*_{i-1}\mathrm{A}_i\mathrm{Cap}_{i+1}$ represents the fragment of the $i$th residue $\mathrm{A}_i$ capped with a left cap $\mathrm{Cap}^*_{i-1}$ and a right cap $\mathrm{Cap}_{i+1}$, and $\mathrm{Cap}^*_i\mathrm{Cap}_{i+1}$ stands for the $i$th concap. $\tilde{E}_i$ and $\tilde{E}_j$ denote the energy of the residue $i$ and $j$ capped with hydrogen atoms on both sides, respectively. $\tilde{E}_{ij}$ is the energy of the dimer consisting of residues $i$ and $j$. $q_{m(k)}$ represents the point charge of the $m$th atom in the fragment $k$. In this study, AMBER94[73] charge model is adopted. In addition, $k = \mathrm{(H)Cap}^*_{i-1} - \mathrm{NH}$ and $l = \mathrm{CO} - \mathrm{A}_{i+2}\mathrm{A}_{i+3}\cdots\mathrm{A}_N$, $(i = 2, 3,$

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

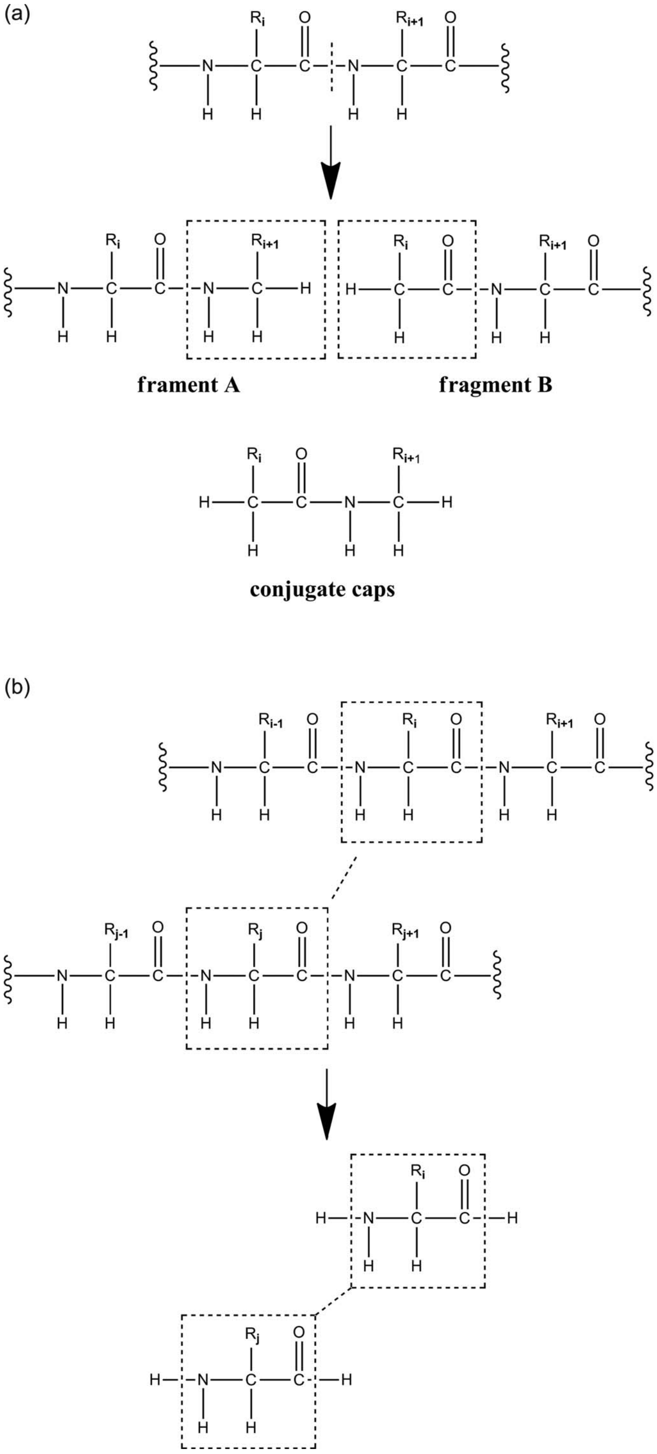

FIG. 1. (a) MFCC scheme in which the peptide bond is cut and the fragments are capped with $\mathrm{Cap}_{i+1}$ and its conjugate $\mathrm{Cap}_i^*$, where subscript $i$ denotes the $i$th amino acid in the given protein. The concap is defined as the fused molecular species of $\mathrm{Cap}_i^* - \mathrm{Cap}_{i+1}$. (b) The generalized concap (Gconcap) scheme and the atomic structure of the Gconcap.

$\ldots, N-2$), where $k$ and $l$ represent all the pairs whose interaction energies have been doubly counted in the MFCC calculation of $\sum_{i=2}^{N-1} \tilde{E}(\mathrm{Cap}_{i-1}^{*}\mathrm{A}_i\mathrm{Cap}_{i+1}) - \sum_{i=2}^{N-2} \tilde{E}(\mathrm{Cap}_i^{*}\mathrm{Cap}_{i+1})$ with the embedding scheme. Furthermore, if the distance of any two atoms between residue $i$ and $j$ is less than or equal to the distance threshold $\lambda$ ($\lambda = 4.0$ Å in this study), the interaction energy between these two residues are calculated by quantum mechanics ($(\tilde{E}_{ij} - \tilde{E}_i - \tilde{E}_j)_{\mathrm{QM}}$); otherwise, the doubly counted interaction energy between distant non-neighboring residue is deducted by charge-charge interactions approximately (the last term in Eq. (1)). The size of the fragment and distance threshold $\lambda$ used in this study have been validated with the 6-31G* basis set in the previous study.[70] Nevertheless, these parameters may vary when the size of the basis set changes, especially for diffuse basis functions. The dependence of the size of the fragment and distance threshold $\lambda$ with respect to the basis sets will be further investigated in future studies. Figure 2 shows the graphic illustration of assembling the overall EE-GMFCC energy for a peptide composed of 4 alanines.

## B. Conductor-like polarized continuum model

The Hamiltonian of a solute molecule in solution is expressed as

$$\mathrm{H} = \mathrm{H}_0 + \mathrm{H}', \tag{2}$$

where $\mathrm{H}_0$ is the gas phase Hamiltonian of the solute molecule and $\mathrm{H}'$ represents the perturbation from the solvent, which is written as

$$\mathrm{H}' = \sum_{\mu,\alpha} \frac{q_\mu Z_\alpha}{|\mathbf{r}_\mu - \mathbf{R}_\alpha|} - \sum_{\mu,i} \frac{q_\mu}{|\mathbf{r}_\mu - \mathbf{r}_i|}, \tag{3}$$

where $q_\mu$ and $\mathbf{r}_\mu$ represent the surface charges and their respective positions, respectively. $Z_\alpha$ and $\mathbf{R}_\alpha$ are the nuclear charges and their corresponding coordinates, and $\mathbf{r}_i$ denotes the position of electron $i$. Normally, the electrostatic solvation energy (G(ele)) is obtained by

$$\begin{aligned}\mathrm{G(ele)} &= \langle\Psi|\mathrm{H}_0|\Psi\rangle - \langle\Psi_0|\mathrm{H}_0|\Psi_0\rangle + \frac{1}{2}\langle\Psi|\mathrm{H}'|\Psi\rangle \\ &= \mathrm{E(wfd)} + \mathrm{G(es)},\end{aligned} \tag{4}$$

where $\Psi_0$ and $\Psi$ represent the solute's wavefunctions in gas phase and in solution, respectively. The wave function distortion energy (E(wfd), or solute polarization energy) is expressed as

$$\mathrm{E(wfd)} = \langle\Psi|\mathrm{H}_0|\Psi\rangle - \langle\Psi_0|\mathrm{H}_0|\Psi_0\rangle \tag{5}$$

and the electrostatic solute-solvent interaction energy is given by

$$\mathrm{G(es)} = \frac{1}{2}\langle\Psi|\mathrm{H}'|\Psi\rangle. \tag{6}$$

In the CPCM method,[68,69,74] a cavity is defined by enveloping spheres centered on the solute atoms and the surface of the cavity is tessellated into mosaics. Each tessera is characterized by its position, area, and normal vector. The continuum solvent is polarized by the solute molecule and induced charges appear on the surface. CPCM provides a simple way to determine the discrete induced surface charge $\mathbf{q}$ by solving the following linear equation:

$$\mathbf{B} + \mathbf{A}\mathbf{q} = \mathbf{0}, \tag{7}$$

where $q_\mu = \sigma_\mu S_\mu$ with $\sigma_\mu$ and $S_\mu$ are, respectively, the surface charge density and discrete surface areas. The matrix

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

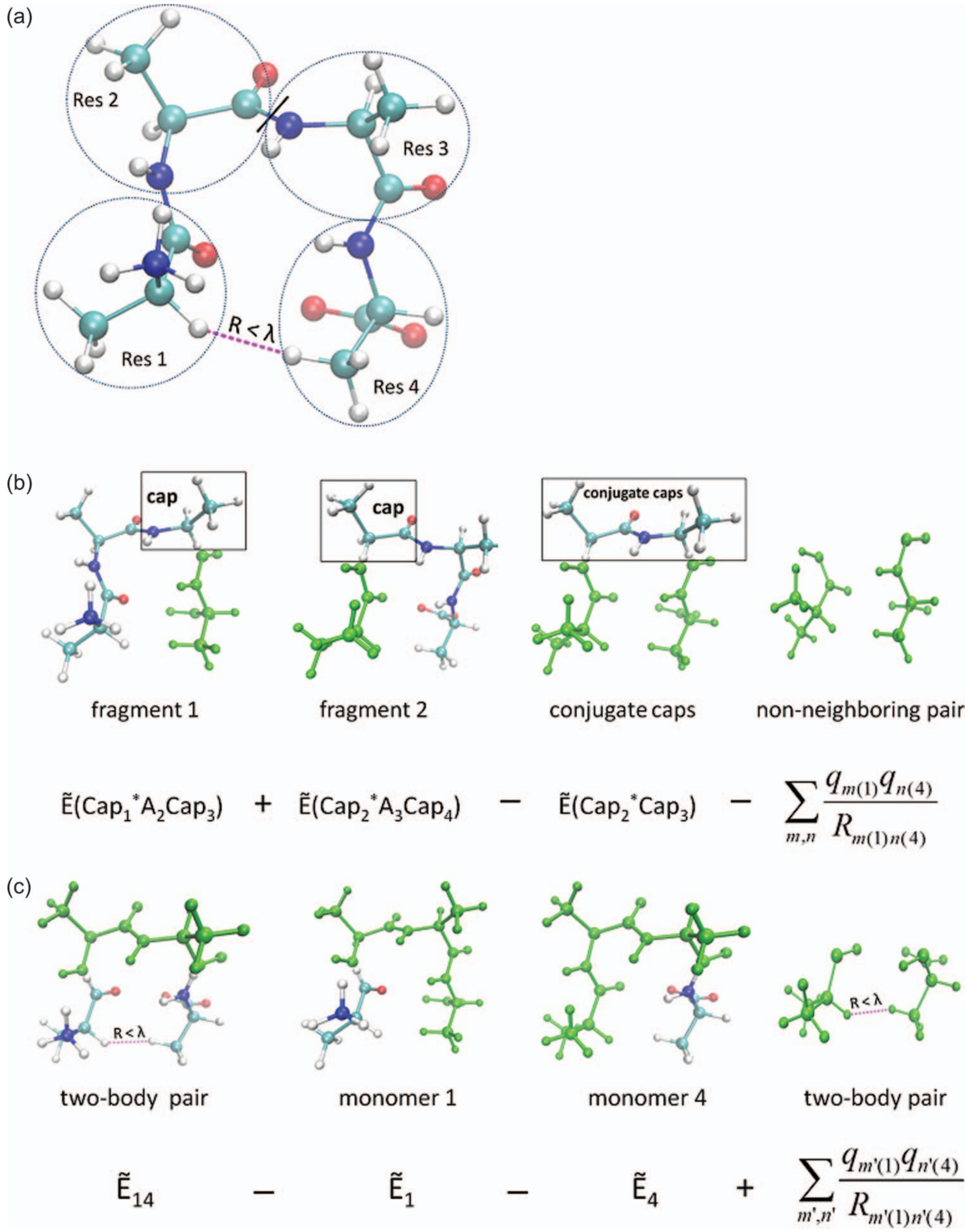
(a)
Res 2
Res 3
Res 1
Res 4
R < λ
(b)
cap
cap
conjugate caps
fragment 1
fragment 2
conjugate caps
non-neighboring pair
$\tilde{E}(\mathrm{Cap}_1{}^*\mathrm{A}_2\mathrm{Cap}_3) + \tilde{E}(\mathrm{Cap}_2{}^*\mathrm{A}_3\mathrm{Cap}_4) - \tilde{E}(\mathrm{Cap}_2{}^*\mathrm{Cap}_3) - \sum_{m,n}\frac{q_{m(1)}q_{n(4)}}{R_{m(1)n(4)}}$
(c)
R < λ
R < λ
two-body pair
monomer 1
monomer 4
two-body pair
$\tilde{E}_{14} - \tilde{E}_1 - \tilde{E}_4 + \sum_{m',n'}\frac{q_{m'(1)}q_{n'(4)}}{R_{m'(1)n'(4)}}$

FIG. 2. (a) The peptide consists of four alanines. It is divided into two fragments by cutting through the peptide bond between residue 2 and residue 3 (as the black line shows). The distance (R) between non-neighboring residues 1 and 4 is less than the threshold λ (as the dashed purple line shows), which means this non-neighboring interaction will be treated with QM method. (b) The atoms colored green are treated as background charges in each fragment QM calculations. The other atoms (not in green) are treated at the QM level. A pair of conjugate caps is shown in the black box. $\tilde{E}$ represents the summation of the self-energy of the fragment and the interaction energy between the fragment and background charges of the remaining system. The third and fourth terms are used to cancel out the energies introduced by the conjugate caps and double counted interactions between residues 1 and 4 in the first two terms, respectively. (c) The two-body interaction energy between residues 1 and 4 is calculated by QM. The fourth term is added to compensate the overly deducted interaction energy between residues 1 and 4 in the first three terms. This compensation partially cancels out the fourth term in (b).

elements are defined by[68,69]

$$\mathbf{A}_{\mu\nu} = \frac{1}{|\mathbf{r}_\mu - \mathbf{r}_\nu|}(1-\delta_{\mu\nu}) + 1.07\sqrt{\frac{4\pi}{S_\mu}}\delta_{\mu\nu} \tag{8}$$

and

$$\mathbf{B}_\mu = \sum_\alpha \frac{Z_\alpha}{|\mathbf{r}_\mu - \mathbf{R}_\alpha|} - \phi(\mathbf{r}_\mu), \tag{9}$$

where $\phi(\mathbf{r}_\mu)$ is the electrostatic potential at surface site $\mu$ created by electrons of the solute molecule, and $\mathbf{r}_\mu$ is the coordinate of the surface site $\mu$. Equation (7) is valid only for conductor with an infinite dielectric constant. The surface charges $q_\mu$ are then corrected for finite dielectric constant by the relation[74]

$$\tilde{q}_\mu = \frac{\varepsilon - 1}{\varepsilon} q_\mu, \tag{10}$$

where $\varepsilon$ is the dielectric constant of the solvent. For small solutes, direct matrix inversion is a common practice in solving Eq. (7). However, for macromolecules with a large

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

number of surface tesserae, iterative methods are required to solve Eq. (7).[75,76] Detailed description on the calculations of energies and other properties of molecules in solution using CPCM method can be found in Ref. 74.

### C. EE-GMFCC-CPCM ansatz

In the EE-GMFCC-CPCM approach, the wave function distortion energy of Eq. (5) can be approximately obtained through

$$\mathrm{E(wfd)} = \sum_{k=1}^{N} \Delta\tilde{E}_k - \sum_{k=1}^{N_c} \Delta\tilde{E}_k^c + \sum_{k=1}^{N_{GC}} \Delta\left(\tilde{E}_k^{ij} - \tilde{E}_k^i - \tilde{E}_k^j\right), \tag{11}$$

where $\tilde{E}$ represents the summation of the self-energy of each fragment and the interaction energy between the fragment and its corresponding background charges of the remaining protein. $N$, $N_C$, and $N_{GC}$ denote the number of capped individual residues, concaps, and Gconcaps, respectively. The wavefuntion distortion energy of individual fragment $\Delta\tilde{E}_k$, conjugate caps $\Delta\tilde{E}_k^c$, and the two-body interaction energy between non-neighboring residues in close contact are given by

$$\Delta\tilde{E}_k = \tilde{E}_k[\Psi_\mathrm{k}] - \tilde{E}_k\left[\Psi_\mathrm{k}^0\right], \tag{12}$$

$$\Delta\tilde{E}_k^c = \tilde{E}_k^c\left[\Psi_\mathrm{k}^\mathrm{c}\right] - \tilde{E}_k^c\left[\Psi_\mathrm{k}^\mathrm{c0}\right], \tag{13}$$

and

$$\begin{aligned}\Delta&\left(\tilde{E}_k^{ij} - \tilde{E}_k^i - \tilde{E}_k^j\right)\\ &= \left(\tilde{E}_k^{ij}\left[\Psi_\mathrm{k}^\mathrm{ij}\right] - \tilde{E}_k^i\left[\Psi_\mathrm{k}^\mathrm{i}\right] - \tilde{E}_k^j\left[\Psi_\mathrm{k}^\mathrm{j}\right]\right)\\ &\quad -\left(\tilde{E}_k^{ij}\left[\Psi_\mathrm{k}^\mathrm{ij0}\right] - \tilde{E}_k^i\left[\Psi_\mathrm{k}^\mathrm{i0}\right] - \tilde{E}_k^j\left[\Psi_\mathrm{k}^\mathrm{j0}\right]\right).\end{aligned} \tag{14}$$

Owing to the fixed charge model used in this study, the doubly counted charge-charge interaction energy (the last term in Eq. (1)) are exactly canceled out between the gas phase and solution phase in the current EE-GMFCC-CPCM calculations.

On the other hand, the electrostatic solute-solvent interaction energy G(es) is given by

$$\begin{aligned}\mathrm{G(es)} &= \frac{1}{2}\sum_{\mu} q_\mu \left[\sum_{\alpha} \frac{Z_\alpha}{|\mathbf{r}_\mu - \mathbf{R}_\alpha|} - \int \frac{\rho(\mathbf{r})}{|\mathbf{r}_\mu - \mathbf{r}|} d\mathbf{r}\right]\\ &= \frac{1}{2}\sum_{\mu} q_\mu \phi(\mathbf{r}_\mu),\end{aligned} \tag{15}$$

where $\phi(\mathbf{r}_\mu)$ is the electrostatic potential on cavity surface site $\mu$ generated by both the nuclei and the electrons of the solute, which can be approximately calculated by adding the two-body correction to the original MFCC contribution as

$$\begin{aligned}\phi(\mathbf{r}_\mu) &= \sum_{k=1}^{N} \phi_k(\mathbf{r}_\mu) - \sum_{k=1}^{N_C} \phi_k^c(\mathbf{r}_\mu)\\ &\quad + \sum_{k=1}^{N_{GC}} \left[\phi_k^{ij}(\mathbf{r}_\mu) - \phi_k^i(\mathbf{r}_\mu) - \phi_k^j(\mathbf{r}_\mu)\right].\end{aligned} \tag{16}$$

In the EE-GMFCC-CPCM approach, the two-body correction was added to both the total energy expression of the solute and the electrostatic potential on the cavity surface, distinguishing this method from the original MFCC-CPCM method. Therefore, the EE-GMFCC-CPCM calculation improves the accuracies of both the wavefunction distortion energy and the electrostatic solute-solvent interaction energy. The QM calculation of each fragment is conducted in the presence of the surface charges. The newly obtained wavefunction is then employed to determine the surface charges of the CPCM equation, Eq. (7). The procedure is repeated in a self-consistent fashion until the final electrostatic solute-solvent interaction energy and the surface charges reach convergence, as in the standard SCRF calculation.

As pointed out in Ref. 65, it is worth noting that the induced charges are defined on the cavity surface of the whole protein, other than that of individual fragment. Therefore, the electrostatic potentials are calculated on the same set of the surface charges for each fragment. Furthermore, the electrostatic solvation energy of a given protein can be readily decomposed into the contributions from each fragment and two-body interactions in the EE-GMFCC-CPCM calculation. In contrast, the standard full system CPCM calculation does not provide the fragment-based decomposition for the solvation energy.

### D. Steps of the EE-GMFCC-CPCM calculation

Similar to our previous study,[65] the cavity that accommodates the protein is generated by the united-atom topological model (UATM)[77] and then tessellated using gepol algorithm[78] implemented in Gaussian 03.[79] The overlap index between two interlocking spheres is 0.89 and the minimum radii for solvent excluded surface (SES) added sphere is 0.2 Å. The average area of tesserae is 0.2 Å$^2$. It is worth noting that, since large proteins usually have some groove areas, the solvent accessible surface (SAS) model may be more suitable than SES. In future biological applications, the SAS model should be utilized. But in this study, the choice of the surface model is not the key purpose. We still utilized the SES model for all calculations. After the cavity is generated, the following steps are as follows:

(1) The EE-GMFCC calculation is performed in gas phase to obtain the initial electrostatic potentials on the surface sites of the cavity. The corresponding induced surface charges are solved based on Eq. (7).
(2) The EE-GMFCC calculation is performed in solution phase represented by the external potential of the induced charges from the previous step. Electrostatic potentials on the surface sites of the cavity are updated.
(3) Update the induced surface charges by solving Eq. (7).
(4) If the root mean square deviations (RMSD) of the induced charges is below $10^{-5}e$ between two successive steps and the variation of $\Delta$G(es) is no larger than 0.5 kcal/mol, the convergence is considered to be reached. Otherwise, step (2) and step (3) are repeated.

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

TABLE I. Electrostatic solvation energies in kcal/mol from the full system CPCM, MFCC-CPCM, and EE-GMFCC-CPCM calculations, respectively, at the HF/6-31G* level.

| Peptide[a] | No. of tesserae | No. of atoms | Energy component[b] | Full system | MFCC-CPCM[c] | EE-GMFCC-CPCM[c] |
|---|---|---|---|---|---|---|
| Pentagly | 3025 | 38 | G(es) | − 143.29 | − 144.55 (−1.26) | − 145.45 (−2.16) |
| | | | E(wfd) | 10.21 | 8.84 (−1.37) | 10.16 (−0.05) |
| | | | G(ele) | − 133.71 | − 135.71 (−2.00) | − 135.29 (−1.58) |
| F1G5Q | 6288 | 72 | G(es) | − 210.31 | − 210.06 (0.25) | − 210.90 (−0.59) |
| | | | E(wfd) | 25.07 | 24.73 (−0.34) | 25.16 (0.09) |
| | | | G(ele) | − 185.23 | − 185.33 (−0.10) | − 185.74 (−0.51) |
| F1P9U | 7691 | 94 | G(es) | − 203.15 | − 201.87 (1.28) | − 204.55 (−1.40) |
| | | | E(wfd) | 22.84 | 21.28 (−1.56) | 22.55 (−0.29) |
| | | | G(ele) | − 180.32 | − 180.59 (−0.27) | − 182.00 (−1.68) |
| Decaala | 6708 | 103 | G(es) | − 255.38 | − 245.38 (10.00) | − 257.14 (−1.76) |
| | | | E(wfd) | 28.59 | 26.84 (−1.75) | 29.44 (0.85) |
| | | | G(ele) | − 226.80 | − 218.54 (8.26) | − 227.70 (−0.90) |
| F1ATP | 8916 | 141 | G(es) | − 315.55 | − 301.45 (14.10) | − 318.91 (−3.36) |
| | | | E(wfd) | 44.15 | 36.45 (−7.70) | 45.07 (0.92) |
| | | | G(ele) | − 271.39 | − 265.00 (6.39) | − 273.84 (−2.45) |
| F1R1W | 12 056 | 168 | G(es) | − 829.97 | − 820.93 (9.04) | − 829.97 (0.00) |
| | | | E(wfd) | 56.71 | 49.43 (−7.28) | 54.23 (−2.48) |
| | | | G(ele) | − 773.26 | − 771.50 (1.76) | − 775.74 (−2.48) |
| F2B4J | 10 508 | 171 | G(es) | − 225.49 | − 221.22 (4.27) | − 226.78 (−1.29) |
| | | | E(wfd) | 26.60 | 20.98 (−5.62) | 27.18 (0.58) |
| | | | G(ele) | − 198.89 | − 200.04 (−1.15) | − 199.60 (−0.71) |
| RMSD[d] | | | G(es) | | 7.57 | 1.81 |
| | | | E(wfd) | | 4.65 | 1.08 |
| | | | G(ele) | | 4.09 | 1.65 |
| MUE[e] | | | G(es) | | 5.74 | 1.51 |
| | | | E(wfd) | | 3.66 | 0.75 |
| | | | G(ele) | | 2.85 | 1.47 |

[a] Pentagly is in an extended structure. F1G5Q: chains P in pdb id 1G5Q, which has two hydrogen bonds. F1P9U: chain G in pdb id 1P9U, which has an extended structure. Decaala is an $\alpha$-helix composed of ten alanines. F1ATP: residues sequence 5–14 of chain 1 in pdb id 1ATP, which adopts an $\alpha$-helix structure. F2B4J: residues 75–86 of chain B in pdb id 2B4J, which is in a $\beta$ strand conformation. F1RIW: chain D in pdb id 1RIW. All peptides have charged termini.

[b] G(ele) = G(es) + E(wfd). G(ele) is the total electrostatic solvation energy. G(es) is the polarized solute-solvent reaction field energy defined in Eq. (6) and E(wfd) is the wavefuntion distortion energy of the solute defined in Eq. (5).

[c] Data in parentheses represent the errors with respect to the full system calculations.

[d] RMSD: root-mean-squared deviation (with respect to the full system results).

[e] MUE: mean unsigned error.

In this study, AMBER94 charge model[73] is employed as the background charges for solute and the dielectric constant for solvent is set to 78.39. All quantum mechanical calculations are carried out at the HF/6-31G* level using the Gaussian 03 package.[79]

### E. Molecular dynamic (MD) simulation

The initial structures of peptides or proteins used in this study are all extracted from Protein Data Bank (PDB). Necessary relaxations through MD simulations are carried out to obtain conformations for MFCC-CPCM, EE-GMFCC-CPCM, and full system calculations. The protein is placed in a periodic TIP3P water box. The distance between the edges of the water box and the closest atom of the solutes is no less than 12 Å. Counterions are added to neutralize the whole system. The protein is initially restrained and all other molecules were relaxed to remove bad contacts using the steep descent method followed by a conjugated gradient optimization. Then the full system is minimized until the convergence is reached. Subsequently, the entire system is heated to 300 K in 100 ps. The equilibrium of the system is simulated in canonical ensemble with a time step of 2 fs for 500 ps followed by a production run for 2 ns in NTP ensemble at 300 K and 1 atm. The temperature is regulated by the Langevin dynamics[80] with a collision frequency of 1.0 ps$^{-1}$, and the Berendsen's barostat[81] is used to regulate the pressure. All covalent bonds involving hydrogen atoms are constrained by SHAKE algorithm. Particle mesh Ewald (PME)[82] method is used to treat the long-range electrostatic interactions in periodic boundary condition. The nonbonded cutoff is set to 10 Å. The last snapshot is selected for calculation of the electrostatic solvation energy. For protein 2I9M, an extra 2 ns MD simulation is carried out for the study of variations of electrostatic solvation energy over the conformational change. All the MD simulations are carried out using the AMBER11 package.[83]

## III. RESULTS AND DISCUSSION

A set of peptides is studied by the standard full system, MFCC-CPCM, and EE-GMFCC-CPCM methods at the HF/6-31G* level. Detailed descriptions of these peptides and the calculated results are listed in Table I. The same set of peptides has also been investigated in our previous MFCC-

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

CPCM study.[65] Results calculated by MFCC-CPCM method in this study are very close to those in the previous study and the subtle variation originates from small differences in the optimized structures. Comparison of the MFCC-CPCM results with full system calculations shows large deviations in the electrostatic solvation energy for F1ATP and Decaala, whose structures consist of $\alpha$-helix. For some other peptides, such as F1R1W and F2B4J, although the total electrostatic solvation energies (G(ele)) are very close to the full system calculations, but large deviations are observed for their components E(wfd) and G(es). Therefore, the overall agreement in G(ele) between the MFCC-PCM and full system results mainly arises from the error cancellation for these two systems. The deviations of MFCC-CPCM from full system results are due to inexact treatment of close contacts between non-neighboring residues such as hydrogen bonding interactions. In contrast, since the EE-GMFCC-CPCM method takes two-body effects into consideration, the deviations mentioned above are expected to be reduced. One can see from Table I that all the electrostatic solvation energies and their components calculated by EE-GMFCC-CPCM are in good agreement with full system calculations. In addition, we found no clear correlation between the magnitude of the error and the size of the solute molecules. Based on these results, the EE-GFMCC-CPCM method is shown to be capable of providing accurate electronic structures of peptides (both in gas phase and in solution).

Next, we utilized the EE-GMFCC-CPCM, MFCC-CPCM methods, and full system calculations at the HF/6-31G* level, respectively, in the studies of several real proteins, which have tertiary structures composed of $\beta$-sheet, $\alpha$-helix, or a mixture of them (see Figure 3). The calculated electrostatic solvation energy G(ele) and its components (G(es), E(wfd)) are listed in Table II. Except for 2XL1, the EE-GMFCC-CPCM calculated electrostatic solvation energies are in excellent agreement with the corresponding full system calculations with all deviations less than 0.01 hartree. For G(ele), G(es), and E(wfd), RMSDs of the EE-GMFCC-CPCM calculations from full system results are 3.93, 2.22, and 2.70 kcal/mol, respectively, which are significantly reduced from 11.26, 7.40, and 13.26 kcal/mol in MFCC-CPCM calculations. Generally speaking, the EE-GMFCC-CPCM method efficiently gives accurate electrostatic solvation energy for large proteins. Despite the approximation of the fragmentation method, the deviations of current EE-GMFCC-CPCM calculations from full system results may also arise from the usage of nonpolarizable AMBER94 atomic charges for the solute. In principle, owing to the electronic perturbation from the induced charges on the surface of molecular cavity, the atomic charges of proteins in solvent should be different from those derived in gas phase. Therefore, we also employed the polarized protein-specific charges (PPCs) as the background charges. The procedure to derive the PPC is described in Ref. 67. Since the PPCs for a given protein in gas phase and in solution are different, the wavefunction distortion energy should also include the difference of the charge-charge interactions (the last term in Eq. (1)) from gas phase to solution. Table S1 in the supplementary material[88] shows the calculated electrostatic solvation energies using the EE-GMCC-CPCM method with the PPC. Similar to the previous EE-GMFCC calculations on the total energies of proteins in gas phase,[70] the overall error of electrostatic solvation energy using PPC with respect to the full system results is slightly larger than that using the fixed AMBER94 charge model. Failure of the PPC may due to the inconsistency between the strategy in charge fitting, which only considers one-body terms, and the subsequent energy calculations with two-body terms. It also shows that using the AMBER94 charge model to approximate the electrostatic potential of the protein is adequate for the electrostatic solvation energy calculation.

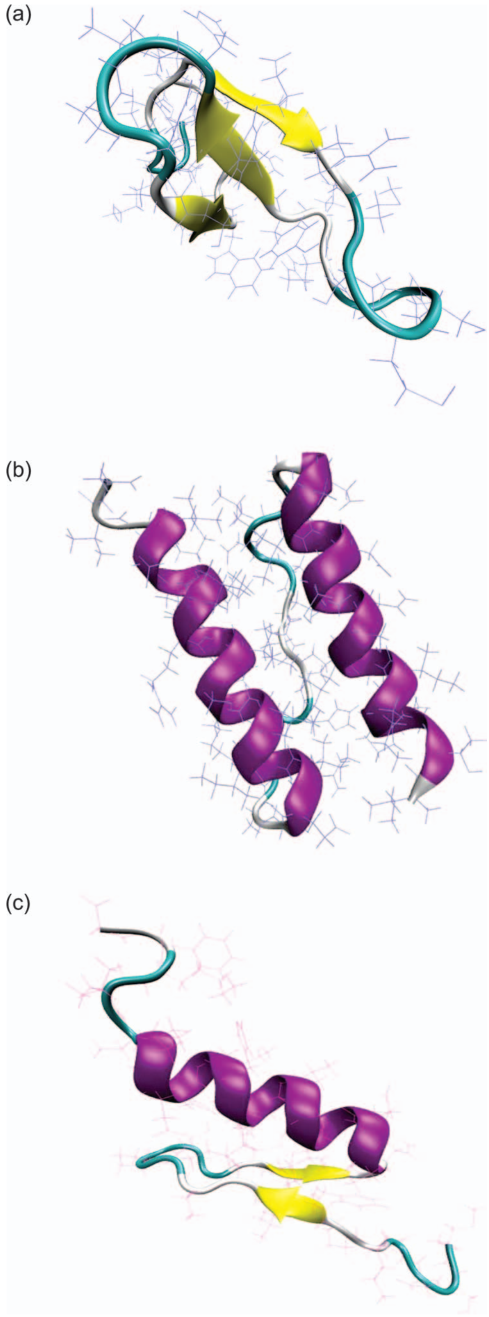

FIG. 3. Three representative three-dimensional protein structures studied in this study. PDB IDs are 2KCF (a), 2Y0Q (b), and 1BHI (c), respectively.

During folding process, protein undergoes large conformational change towards a compact structure. The interplay

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

TABLE II. Electrostatic solvation energies in kcal/mol from the full system CPCM, MFCC-CPCM, and EE-GMFCC-CPCM calculations, respectively, at the HF/6-31G* level.

| Protein (pdb id) | No. of tesserae | No. of atoms | Energy component | Full system | MFCC-CPCM[a] | EE-GMFCC-CPCM[a] |
|---|---|---|---|---|---|---|
| 1LE1 | 10 756 | 218 | G(es) | – 188.37 | – 189.84 (-1.47) | – 189.40 (−1.03) |
| | | | E(wfd) | 21.64 | 14.64 (−7.00) | 18.26 (−3.38) |
| | | | G(ele) | – 166.73 | – 175.20 (−8.47) | – 171.14 (−4.41) |
| 2I9M | 14 360 | 246 | G(es) | – 264.63 | – 254.34 (10.29) | – 266.47 (−1.84) |
| | | | E(wfd) | 27.48 | 24.75 (−2.73) | 28.18 (0.70) |
| | | | G(ele) | – 237.15 | – 229.59 (7.56) | – 238.29 (−1.14) |
| 2GB1 | 14 013 | 247 | G(es) | – 482.72 | – 483.50 (−0.78) | – 483.84 (−1.12) |
| | | | E(wfd) | 43.02 | 34.36 (−8.66) | 40.90 (−2.12) |
| | | | G(ele) | – 439.70 | – 449.14 (−9.44) | – 442.94 (−3.24) |
| 1L2Y | 16 642 | 304 | G(es) | – 372.86 | – 363.92 (8.94) | – 374.33 (−1.47) |
| | | | E(wfd) | 48.08 | 42.74 (−5.34) | 45.45 (−2.63) |
| | | | G(ele) | – 324.78 | – 321.18 (3.60) | – 328.88 (−4.10) |
| 1WN8 | 19 701 | 354 | G(es) | – 285.43 | – 278.01 (7.42) | – 286.70 (−1.27) |
| | | | E(wfd) | 27.44 | 27.42 (−0.02) | 27.71 (0.27) |
| | | | G(ele) | – 257.99 | – 250.59 (7.40) | – 258.99 (−1.00) |
| 2XL1 | 12 369 | 243 | G(es) | – 203.88 | – 196.32 (7.56) | – 205.70 (−1.82) |
| | | | E(wfd) | 26.87 | 20.83 (−6.04) | 20.80 (−6.07) |
| | | | G(e | – 177.02 | – 175.49 (1.53) | – 184.90 (−7.88) |
| 1VTP | 27 004 | 396 | G(es) | – 882.48 | – 873.10 (9.38) | – 882.71 (−0.23) |
| | | | E(wfd) | 100.86 | 76.00 (−24.86) | 96.46 (−4.40) |
| | | | G(ele) | – 781.62 | – 797.10 (−15.48) | – 786.24 (−4.62) |
| 1BBA | 36 259 | 582 | G(es) | – 698.31 | – 703.77 (−5.46) | – 695.68 (2.63) |
| | | | E(wfd) | 75.06 | 53.08 (−21.98) | 73.57 (−1.49) |
| | | | G(ele) | – 623.25 | – 650.69 (−27.44) | – 622.11 (1.14) |
| 2KCF | 29 995 | 576 | G(es) | – 745.74 | – 737.88 (7.86) | – 745.67 (0.07) |
| | | | E(wfd) | 60.27 | 51.22 (−9.05) | 59.27 (−1.00) |
| | | | G(ele) | – 685.46 | – 686.66 (−1.20) | – 686.40 (−0.94) |
| 2F21 | 32 205 | 603 | G(es) | – 711.35 | – 705.08 (6.27) | – 705.98 (5.37) |
| | | | E(wfd) | 86.56 | 76.45 (−10.11) | 86.19 (−0.37) |
| | | | G(ele) | – 624.79 | – 628.63 (−3.84) | – 619.79 (5.00) |
| 1BHI | 32 358 | 591 | G(es) | – 597.68 | – 595.28 (2.40) | – 595.40 (2.28) |
| | | | E(wfd) | 69.58 | 55.31 (−14.27) | 69.79 (0.21) |
| | | | G(ele) | – 528.10 | – 539.97 (−11.87) | – 525.61 (2.49) |
| 2Y0Q | 32 003 | 803 | G(es) | – 571.71 | – 560.20 (11.51) | – 569.43 (2.28) |
| | | | E(wfd) | 68.24 | 47.61 (−20.63) | 70.39 (2.15) |
| | | | G(ele) | – 503.47 | – 512.59 (−9.12) | – 499.04 (4.43) |
| RMSD | | | G(es) | | 7.40 | 2.22 |
| | | | E(wfd) | | 13.26 | 2.70 |
| | | | G(ele) | | 11.26 | 3.93 |
| MUE | | | G(es) | | 6.61 | 1.78 |
| | | | E(wfd) | | 10.89 | 2.07 |
| | | | G(ele) | | 8.91 | 3.36 |

[a] Data in parentheses represent the errors with respect to the full system calculations.

between the protein and solvent has a great impact on the folding pathway and other thermodynamic properties of the protein. The relative electrostatic solvation energies of different conformations of protein 2I9M are calculated using fragmentation methods and compared with full system results. Protein 2I9M is a prototype of $\alpha$-helical polypeptide that has attracted some theoretical folding studies.[84,85] Because the AMBER99SB force field is considered to disfavor the $\alpha$-helical structure,[86] the native conformation of 2I9M is not stable and unfolding occurs during MD simulation. As a result, it is worth studying the free energy profile of protein 2I9M in solvent at *ab initio* level.[57] Nineteen conformations are selected at an interval of 100 ps from 2 ns MD simulation. The energies of G(ele), G(es), and E(wfd) calculated for the 19 conformations are shown in Figure 4. The total electrostatic solvation energy undergoes large fluctuation between −470 and −300 kcal/mol. Results from the EE-GMFCC-CPCM calculations are in excellent agreement with full system calculations. The RMSDs of G(ele), G(es), and E(wfd) are 1.14, 1.99, and 0.93 kcal/mol, respectively (see Table III). In contrast, much larger deviations are observed for MFCC-CPCM calculations. The corresponding RMSDs of G(ele), G(es), and E(wfd) are 4.64, 9.70, and 6.99 kcal/mol, respectively. Again, the EE-GMFCC-CPCM approach shows

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

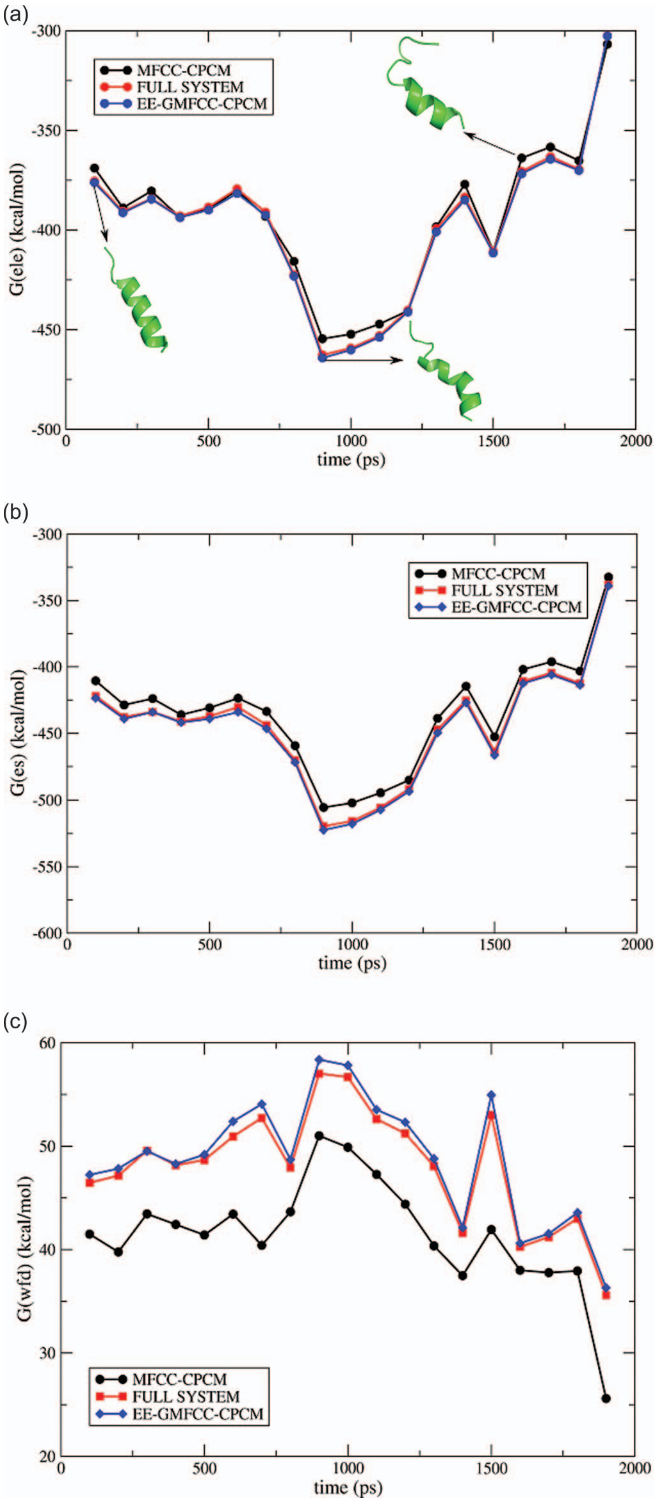

FIG. 4. Variations of G(ele) (a) and its components G(es) (b) and E(wfd) (c) over 19 different conformations selected from 2 ns MD simulation for the protein 2I9M. The calculations are performed using the MFCC-CPCM, EE-GMFCC-CPCM, and standard full system HF/6-31G* methods, respectively.

TABLE III. The statistical deviations of the electrostatic solvation energy (in kcal/mol) calculated by the MFCC-CPCM and EE-GMFCC-CPCM methods with respect to the full system calculations on 19 selected MD snapshots of protein 2I9M at the HF/6-31G* level. MUE and RMSD denote the mean unsigned error and root-mean-squared deviation, respectively.

| | MFCC-CPCM | | | EE-GMFCC-CPCM | | |
|---|---|---|---|---|---|---|
| | G(es) | E(wfd) | G(ele) | G(es) | E(wfd) | G(ele) |
| MUE | 9.37 | 6.54 | 3.79 | 1.82 | 0.80 | 1.02 |
| RMSD | 9.70 | 6.99 | 4.64 | 1.99 | 0.93 | 1.14 |

significant improvement over the MFCC-CPCM results, which underscores that the inclusion of two-body QM interaction is crucial for accurate calculation of protein energies both in gas phase and in solvent.

We further applied the EE-GMFCC-CPCM approach to calculate the electrostatic solvation energies for a number of large globular proteins. Since full system calculations are not feasible for these systems, we compare the results with those from divide-and-conquer PM3 calculations with the Poisson-Boltzmann solvation model (D&C-PB, CM2 charge model).[22] As shown in Figure 5, three largest globular proteins (PDB ID: 2KBO, 4B2F, and 1SBT) utilized in this study

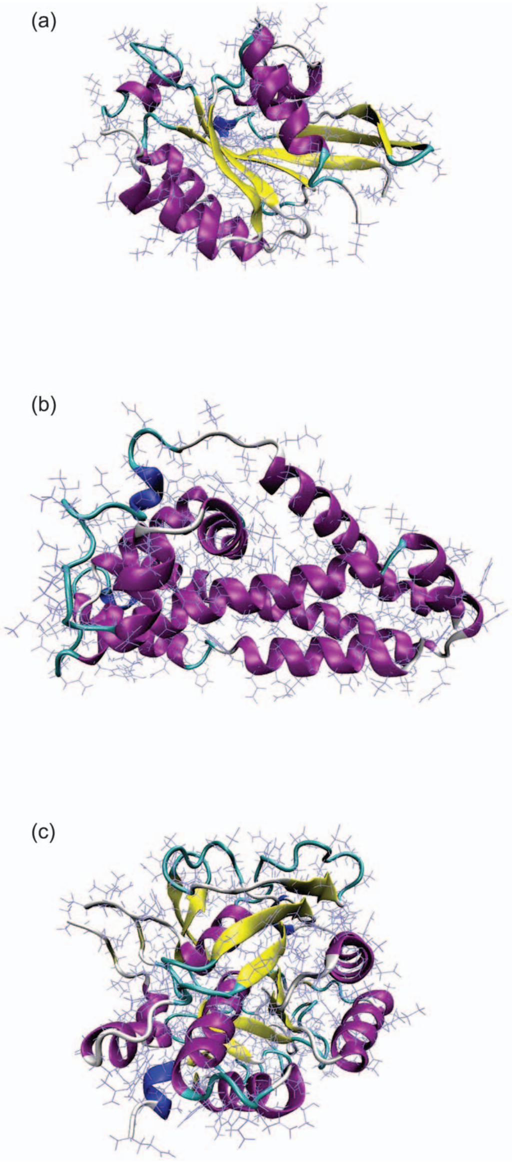

FIG. 5. Proteins with PDB id of 2KBO (a), 4B2F (b), and 1SBT (c) used for MFCC-CPCM, EE-GMFCC-CPCM(HF/6-31G*), and D&C-PB(PM3) calculations.

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

TABLE IV. Computed electrostatic solvation energies (in kcal/mol) from the MFCC-CPCM, EE-GMFCC-CPCM calculations at the HF/6-31G* level, and D&C-PB methods, respectively.

| Protein (pdb id) | No. of tesserae | No. of Atom | Energy component | MFCC-CPCM | EE-GMFCC-CPCM | D&C-PB[a] |
|---|---|---|---|---|---|---|
| 2JOF | 15 562 | 284 | G(es) | − 341.29 | − 347.23 | − 339.87 |
| | | | E(wfd) | 40.36 | 39.05 | 26.51 |
| | | | G(ele) | − 300.93 | − 308.18 | − 313.36 |
| 1AMC | 29 977 | 438 | G(es) | − 861.48 | − 866.70 | − 907.65 |
| | | | E(wfd) | 83.52 | 92.43 | 68.62 |
| | | | G(ele) | − 777.96 | − 774.27 | − 839.03 |
| 2F4K | 32 708 | 538 | G(es) | − 610.76 | − 615.54 | − 664.85 |
| | | | E(wfd) | 60.61 | 72.01 | 54.81 |
| | | | G(ele) | − 550.15 | − 543.53 | − 610.04 |
| 2WXC | 43 901 | 710 | G(es) | − 971.71 | − 992.56 | − 1000.26 |
| | | | E(wfd) | 96.38 | 124.44 | 89.16 |
| | | | G(ele) | − 875.33 | − 868.12 | − 911.10 |
| 1PRB | 45 944 | 851 | G(es) | − 1033.67 | − 1047.35 | − 1132.73 |
| | | | E(wfd) | 107.36 | 127.84 | 96.36 |
| | | | G(ele) | − 926.31 | − 919.51 | − 1036.37 |
| 2A3D | 53 052 | 1140 | G(es) | − 1458.63 | − 1485.28 | − 1571.31 |
| | | | E(wfd) | 133.29 | 171.52 | 129.46 |
| | | | G(ele) | − 1325.34 | − 1313.76 | − 1441.85 |
| 1CDN | 53 209 | 1195 | G(es) | − 1943.24 | − 1920.49 | − 2084.84 |
| | | | E(wfd) | 106.84 | 189.80 | 147.82 |
| | | | G(ele) | − 1836.40 | − 1730.69 | − 1937.02 |
| 2M71 | 71 293 | 1633 | G(es) | − 1897.16 | − 1924.96 | − 2057.41 |
| | | | E(wfd) | 173.55 | 215.52 | 169.23 |
| | | | G(ele) | − 1723.61 | − 1709.44 | − 1888.18 |
| 2M5V | 83 752 | 1661 | G(es) | − 1876.21 | − 1897.14 | − 2068.17 |
| | | | E(wfd) | 155.70 | 218.37 | 167.84 |
| | | | G(ele) | − 1720.51 | − 1678.77 | − 1900.33 |
| 3ZCO | 81 576 | 2080 | G(es) | − 1626.02 | − 1650.22 | − 1757.59 |
| | | | E(wfd) | 161.89 | 223.95 | 167.81 |
| | | | G(ele) | − 1464.15 | − 1426.27 | − 1589.78 |
| 4K8J | 88 319 | 2210 | G(es) | − 2503.91 | − 2518.42 | − 2743.69 |
| | | | E(wfd) | 179.00 | 227.86 | 172.72 |
| | | | G(ele) | − 2324.91 | − 2290.47 | − 2570.97 |
| 2L7N | 102 438 | 2449 | G(es) | − 1943.90 | − 1946.19 | − 2129.96 |
| | | | E(wfd) | 160.58 | 223.15 | 169.86 |
| | | | G(ele) | − 1783.32 | − 1723.04 | − 1960.10 |
| 2KBO | 128 781 | 2790 | G(es) | − 2693.57 | − 2703.52 | − 2926.41 |
| | | | E(wfd) | 258.16 | 318.65 | 244.72 |
| | | | G(ele) | − 2435.41 | − 2384.87 | − 2681.69 |
| 4B2F | 126 408 | 3347 | G(es) | − 2457.90 | − 2467.62 | − 2707.01 |
| | | | E(wfd) | 197.94 | 313.54 | 242.67 |
| | | | G(ele) | − 2259.96 | − 2154.08 | − 2464.34 |
| 1SBT | 121 536 | 3837 | G(es) | − 2177.95 | − 2212.76 | − 2244.20 |
| | | | E(wfd) | 211.90 | 311.97 | 227.15 |
| | | | G(ele) | − 1966.05 | − 1900.79 | − 2017.05 |

[a] The DCQTP program[22] developed by Professor Kenneth M. Merz, Jr.'s group is used. Results are calculated at the PM3 level with CM2 charges.

are all composed of more than 2500 atoms. All the results are given in Table IV. Figure 6 shows a direct comparison between the MFCC-CPCM and EE-GMFCC-CPCM with D&C-PB results. Similar to the previous study,[65] both MFCC-CPCM and EE-GMFCC-CPCM are linearly correlated with D&C-PB calculations for G(ele) and G(es) with correlation coefficients ($R^2$) over 0.996, which may not be surprising due to the large span of the energy range. However, for the wave function distortion term E(wfd), the EE-GMFCC-CPCM gives a better correlation coefficient of 0.996 with D&C-PB compared to 0.927 obtained using the MFCC-CPCM method. As discussed in the above results, E(wfd) calculated by EE-GMFCC-CPCM is in better agreement with the full system calculation than that by MFCC-CPCM, indicating that both EE-GMFCC-CPCM and D&C-PB provide qualitatively correct description of the wavefunction distortion energy of

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

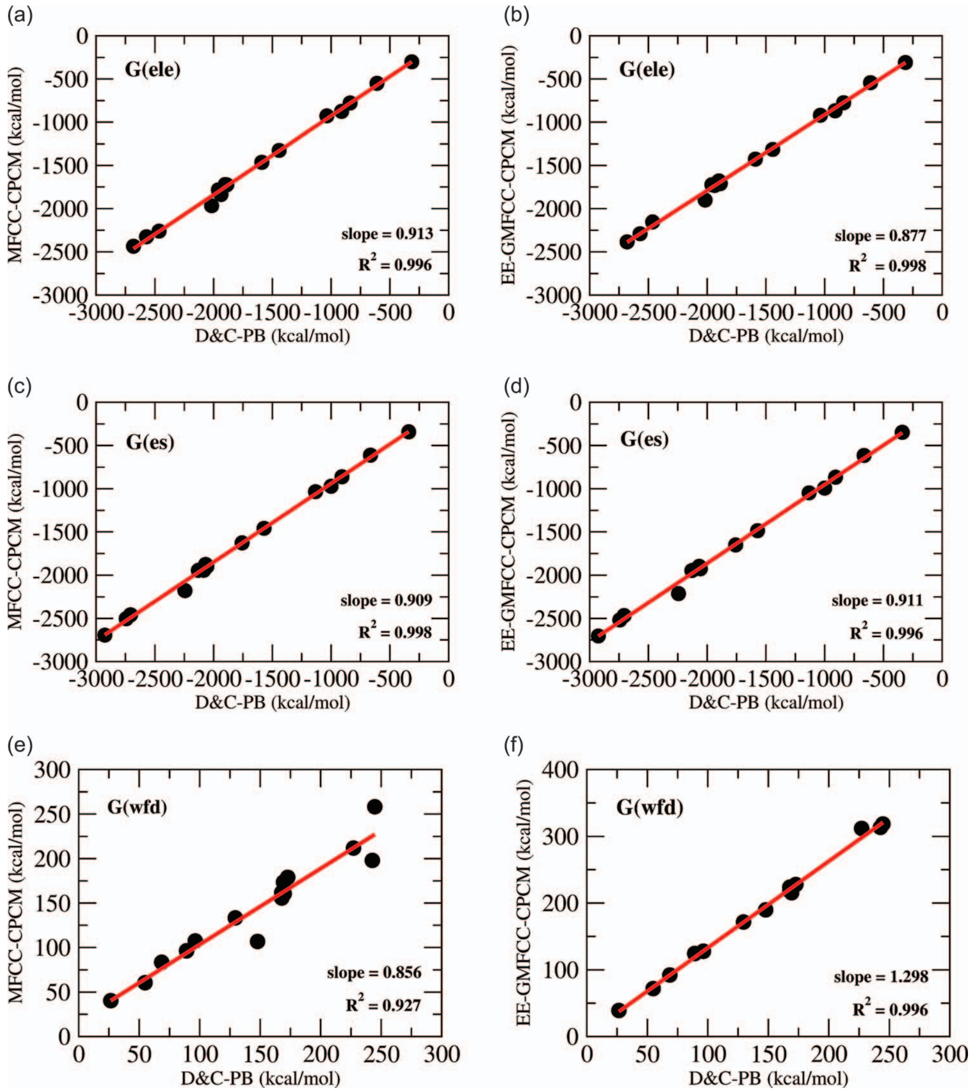

FIG. 6. Correlations between MFCC-CPCM(HF/6-31G*) and D&C-PB(PM3) results for G(ele) (a), G(es) (c), and E(wfd) (e), and correlations between EE-GMFCC-CPCM(HF/6-31G*) and D&C-PB(PM3) results for G(ele) (b), G(es) (d), and E(wfd) (f), respectively.

solute in solvent. Furthermore, as shown in Figures 6(c) and 6(d), the slopes of MFCC-CPCM and EE-GMFCC-CPCM for G(es) are 0.909 and 0.911, respectively, which reveals that the relative solute-solvent interaction energy between different sizes of proteins calculated by QM fragmentation method is smaller than the D&C-PB method. The same trend is observed in Figures 6(a) and 6(b), since the electrostatic solvation energy is dominated by the solute-solvent interaction energy. On the contrary, as shown in Figure 6(f), the slope of EE-GMFCC-CPCM for wavefunction distortion energy is 1.298, indicating that the EE-GMFCC-CPCM approach gives larger relative E(wfd) between different sizes of proteins than the D&C-PB method. But for MFCC-CPCM, the slope of the fitted line for wavefunction distortion energy is 0.856, showing that the relative E(wfd) calculated by MFCC-CPCM is smaller than the D&C-PB method. After the two-body interaction for short-range non-neighboring residues are included in the EE-MFCC-CPCM method, the trend is reversed. Therefore, the two-body interaction energy correction is indispensible in the fragmentation QM method to capture correct E(wfd), resulting in a more accurate calculation of the electrostatic solvation energy for proteins.

We also studied the computational efficiency of the EE-GMFCC-CPCM method. As has been pointed out in a previous study,[70] the total computational time ($T$) for the EE-GMFCC method can be approximately expressed as

$$T \approx \left[\overline{\alpha_3} + \overline{\alpha_2}\left(1 + \frac{\overline{P}}{2}\right) + \overline{\alpha_1}\right] N,$$

where $\overline{P}$ is the average number of amino acids within the distance threshold $\lambda$ from each residue. $\overline{\alpha_3}$, $\overline{\alpha_2}$, and $\overline{\alpha_1}$ are the average computational time for capped fragment $Cap_{i-1}^{*}A_iCap_{i+1}$, concap $Cap_i^{*}Cap_{i+1}$ or Gconcap, and the single residue, respectively. $N$ stands for the total number of residues. In this study, the total EE-GMFCC-CPCM computational time can be obtained by multiplying $T$ by the number of iterative cycles. A comparison of the CPU time on the 8-core Intel Xeon E5620 2.4 GHz processor using the full system HF and EE-GMFCC-CPCM approaches based on the proteins selected from Table II is shown in Figure 7. As expected, the

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

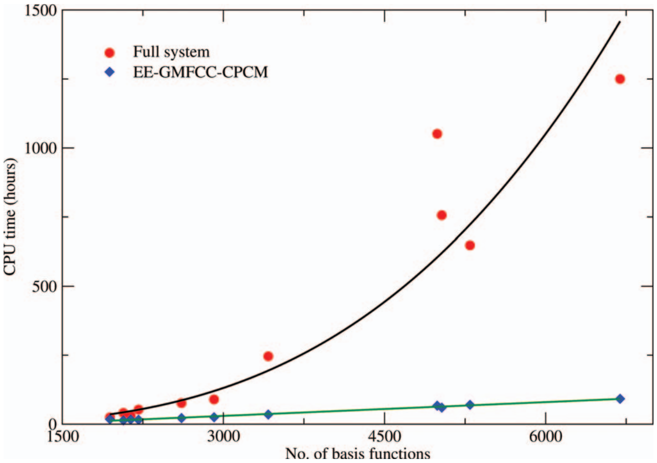

FIG. 7. CPU time for the full system and EE-GMFCC-CPCM calculations as a function of the number of basis functions at the HF/6-31G* level.

computational time scale for the EE-GMFCC-CPCM calculation is O($N$) with a low prefactor, in contrast to O($N^3$) for the traditional HF calculation of the entire system.

## IV. SUMMARY

The EE-GMFCC-CPCM approach is an efficient linear-scaling QM method to treat biological macromolecules in solution. Compared with original MFCC-CPCM method, the crucial aspect of EE-GMFCC is to use QM calculation to deal with short-range non-neighboring interaction. This rigorous treatment of the two-body effect is necessary because this near field interaction cannot be merely treated as simple Coulomb interaction. Other terms such as dispersion, exchange and charge transfer effects require quantum mechanical representation.

In this study, the EE-GMFCC-CPCM approach is applied to calculate the electrostatic solvation energies for a variety of peptides and proteins with up to 3837 atoms. In the previous MFCC-CPCM study, small deviations from full system calculations appear for peptides which are composed of $\alpha$-helical structure. The deviations from the full system results are significantly reduced when the short-range two-body interactions are included in the EE-GMFCC-CPCM approach. For 12 realistic three-dimensional proteins, the electrostatic solvation energy calculated using the EE-GMFCC-CPCM method gives much better agreement with the full system QM calculation than MFCC-CPCM. RMSDs of the electrostatic solvation energy for EE-GMFCC-CPCM and MFCC-CPCM are 3.93 and 11.26 kcal/mol, respectively, with respect to the full system results.

We also compared the results from EE-GMFCC-CPCM(HF/6-31G*) calculation with those from D&C-PB(PM3). The electrostatic solvation energies of proteins (G(ele)) and their components (the polarized solute-solvent reaction field energy G(es) and wavefunction distortion energy E(wfd)) from these two calculations are all highly correlated. However, the relative E(wfd) calculated by EE-GMFCC-CPCM between different sizes of proteins are larger than those from D&C-PB. On the other hand, both the relative G(ele) and G(es) energies given by EE-GMFCC-CPCM are smaller than D&C-PB.

Owing to linear scaling and computational efficiency, the EE-GMFCC-CPCM approach is an efficient method for quantum study of proteins in solution. EE-GMFCC-CPCM can also be applied with density functional theory or other high-level electron correlation methods.[70] Moreover, other implicit solvent models[8] like SMD,[9] MST[87] can also be combined with the EE-GMFCC approach to further improve the accuracy of theoretical calculation on the electrostatic solvation energy of proteins. We believe that the current method will help quantify the protein free energy in solution at *ab initio* levels.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 10974054, 20933002, 21173082, and 21303057) and the Shanghai PuJiang Program (Grant No. 09PJ1404000). X.H. is also supported by the Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP, Grant No. 20130076120019) and the Fundamental Research Funds for the Central Universities. We also thank the Supercomputer Center of East China Normal University for providing us computational time.

[1] V. Limongelli, L. Marinelli, S. Cosconati, C. La Motta, S. Sartini, L. Mugnaini, F. Da Settimo, E. Novellino, and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **109**, 1467 (2012).
[2] J. H. Tian and A. E. Garcia, J. Am. Chem. Soc. **133**, 15157 (2011).
[3] R. Ferreira, A. Marchand, and V. Gabelica, Methods **57**, 56 (2012).
[4] Y. X. Lu, H. Y. Li, X. Zhu, H. L. Liu, and W. L. Zhu, Int. J. Quantum Chem. **112**, 1421 (2012).
[5] J. Tomasi and M. Persico, Chem. Rev. **94**, 2027 (1994).
[6] C. J. Cramer and D. G. Truhlar, Chem. Rev. **99**, 2161 (1999).
[7] J. Tomasi, B. Mennucci, and R. Cammi, Chem. Rev. **105**, 2999 (2005).
[8] M. Kolar, J. Fanfrlik, M. Lepsik, F. Forti, F. J. Luque, and P. Hobza, J. Phys. Chem. B **117**, 5950 (2013).
[9] A. V. Marenich, C. J. Cramer, and D. G. Truhlar, J. Phys. Chem. B **113**, 6378 (2009).
[10] R. Cammi and J. Tomasi, J. Comput. Chem. **16**, 1449 (1995).
[11] C. Amovilli, V. Barone, R. Cammi, E. Cances, M. Cossi, B. Mennucci, C. S. Pomelli, and J. Tomasi, in *Advances in Quantum Chemistry*, Quantum Systems in Chemistry and Physics, Part II Vol. 32, edited by P. O. Lowdin, J. R. Sabin, M. C. Zerner, E. Brandas, S. Wilson, J. Maruani, Y. G. Smeyers, P. J. Grout, and R. McWeeny (Academic Press Inc, San Diego, 1999), p. 227.
[12] E. Cances, B. Mennucci, and J. Tomasi, J. Chem. Phys. **107**, 3032 (1997).
[13] D. M. York and M. Karplus, J. Phys. Chem. A **103**, 11060 (1999).
[14] G. Scalmani and M. J. Frisch, J. Chem. Phys. **132**, 114110 (2010).
[15] A. W. Lange and J. M. Herbert, J. Phys. Chem. Lett. **1**, 556 (2010).
[16] J. B. Foresman, T. A. Keith, K. B. Wiberg, J. Snoonian, and M. J. Frisch, J. Phys. Chem. **100**, 16098 (1996).
[17] J. G. Kirkwood, J. Chem. Phys. **2**, 351 (1934).
[18] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, J. Am. Chem. Soc. **112**, 6127 (1990).
[19] M. Schaefer and M. Karplus, J. Phys. Chem. **100**, 1578 (1996).
[20] S. R. Edinger, C. Cortis, P. S. Shenkin, and R. A. Friesner, J. Phys. Chem. B **101**, 1190 (1997).
[21] D. J. Tannor, B. Marten, R. Murphy, R. A. Freisner, D. Sitkoff, A. Nicholls, M. Ringnalda, W. A. I. Goddardi, and B. Honig, J. Am. Chem. Soc. **116**, 11875 (1994).
[22] V. Gogonea and K. M. Merz, J. Phys. Chem. A **103**, 5171 (1999).
[23] M. E. Davis and J. A. McCammon, Chem. Rev. **90**, 509 (1990).
[24] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi, Chem. Phys. Lett. **313**, 701 (1999).
[25] D. G. Fedorov and K. Kitaura, J. Phys. Chem. A **111**, 6904 (2007).

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10

[26] T. Nakano, T. Kaminuma, T. Sato, K. Fukuzawa, Y. Akiyama, M. Uebayasi, and K. Kitaura, Chem. Phys. Lett. **351**, 475 (2002).

[27] K. Babu, V. Ganesh, S. R. Gadre, and N. E. Ghermani, Theor. Chem. Acc. **111**, 255 (2004).

[28] V. Ganesh, R. K. Dongare, P. Balanarayan, and S. R. Gadre, J. Chem. Phys. **125**, 104109 (2006).

[29] K. Babu and S. R. Gadre, J. Comput. Chem. **24**, 484 (2003).

[30] M. Isegawa, B. Wang, and D. G. Truhlar, J. Chem. Theory Comput. **9**, 1381 (2013).

[31] M. A. Collins and V. A. Deev, J. Chem. Phys. **125**, 104104 (2006).

[32] V. Deev and M. A. Collins, J. Chem. Phys. **122**, 154102 (2005).

[33] J. M. Mullin, L. B. Roskop, S. R. Pruitt, M. A. Collins, and M. S. Gordon, J. Phys. Chem. A **113**, 10040 (2009).

[34] H. M. Netzloff and M. A. Collins, J. Chem. Phys. **127**, 134113 (2007).

[35] T. E. Exner and P. G. Mezey, J. Phys. Chem. A **106**, 11791 (2002).

[36] T. E. Exner and P. G. Mezey, J. Phys. Chem. A **108**, 4301 (2004).

[37] T. E. Exner and P. G. Mezey, Phys. Chem. Chem. Phys. **7**, 4061 (2005).

[38] E. E. Dahlke and D. G. Truhlar, J. Chem. Theory Comput. **3**, 46 (2007).

[39] E. E. Dahlke and D. G. Truhlar, J. Chem. Theory Comput. **3**, 1342 (2007).

[40] A. Sorkin, E. E. Dahlke, and D. G. Truhlar, J. Chem. Theory Comput. **4**, 683 (2008).

[41] S. Li, W. Li, and T. Fang, J. Am. Chem. Soc. **127**, 7215 (2005).

[42] W. Li, S. Li, and Y. Jiang, J. Phys. Chem. A **111**, 2193 (2007).

[43] S. Hua, W. Hua, and S. Li, J. Phys. Chem. A **114**, 8126 (2010).

[44] D. W. Zhang and J. Z. H. Zhang, J. Chem. Phys. **119**, 3599 (2003).

[45] X. He and J. Z. H. Zhang, J. Chem. Phys. **122**, 031103 (2005).

[46] X. He and J. Z. H. Zhang, J. Chem. Phys. **124**, 184703 (2006).

[47] X. H. Chen, Y. K. Zhang, and J. Z. H. Zhang, J. Chem. Phys. **122**, 184105 (2005).

[48] X. H. Chen and J. Z. H. Zhang, J. Chem. Phys. **125**, 044903 (2006).

[49] Y. Mei, X. He, C. G. Ji, D. W. Zhang, and J. Z. H. Zhang, Prog. Chem. **24**, 1058 (2012).

[50] A. M. Gao, D. W. Zhang, J. Z. H. Zhang, and Y. K. Zhang, Chem. Phys. Lett. **394**, 293 (2004).

[51] M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. V. Slipchenko, Chem. Rev. **112**, 632 (2012).

[52] P. N. Day, R. Pachter, M. S. Gordon, and G. N. Merrill, J. Chem. Phys. **112**, 2063 (2000).

[53] P. Bandyopadhyay, Theor. Chem. Acc. **120**, 307 (2008).

[54] X. He, O. Sode, S. S. Xantheas, and S. Hirata, J. Chem. Phys. **137**, 204505 (2012).

[55] Y. Komeiji, T. Ishida, D. G. Fedorov, and K. Kitaura, J. Comput. Chem. **28**, 1750 (2007).

[56] T. Sawada, D. G. Fedorov, and K. Kitaura, Int. J. Quantum Chem. **109**, 2033 (2009).

[57] X. He, L. Fusti-Molnar, G. L. Cui, and K. M. Merz, J. Phys. Chem. B **113**, 5290 (2009).

[58] X. He and K. M. Merz, Jr., J. Chem. Theory Comput. **6**, 405 (2010).

[59] T. Sawada, D. G. Fedorov, and K. Kitaura, J. Am. Chem. Soc. **132**, 16862 (2010).

[60] K. Takematsu, K. Fukuzawa, K. Omagari, S. Nakajima, K. Nakajima, Y. Mochizuki, T. Nakano, H. Watanabe, and S. Tanaka, J. Phys. Chem. B **113**, 4991 (2009).

[61] L. Huang, L. Massa, and J. Karle, Proc. Natl. Acad. Sci. U.S.A. **104**, 4261 (2007).

[62] Y. Orimoto, F. L. Gu, A. Imamura, and Y. Aoki, J. Chem. Phys. **126**, 215104 (2007).

[63] X. He, Y. Mei, Y. Xiang, D. W. Zhang, and J. Z. H. Zhang, Proteins: Struct., Funct., Bioinf. **61**, 423 (2005).

[64] Y. Mei, X. He, Y. Xiang, D. W. Zhang, and J. Z. H. Zhang, Proteins: Struct., Funct., Bioinf. **59**, 489 (2005).

[65] Y. Mei, C. G. Ji, and J. Z. H. Zhang, J. Chem. Phys. **125**, 094906 (2006).

[66] D. G. Fedorov, K. Kitaura, H. Li, J. H. Jensen, and M. S. Gordon, J. Comput. Chem. **27**, 976 (2006).

[67] C. Ji, Y. Mei, and J. Z. H. Zhang, Biophys. J. **95**, 1080 (2008).

[68] A. Klamt and G. Schuurmann, J. Chem. Soc., Perkin Trans. 2 **1993**, 799.

[69] V. Barone and M. Cossi, J. Phys. Chem. A **102**, 1995 (1998).

[70] X. W. Wang, J. F. Liu, J. Z. H. Zhang, and X. He, J. Phys. Chem. A **117**, 7149 (2013).

[71] N. J. Mayhall and K. Raghavachari, J. Chem. Theory Comput. **8**, 2669 (2012).

[72] H.-A. Le, H.-J. Tan, J. F. Ouyang, and R. P. A. Bettens, J. Chem. Theory Comput. **8**, 469 (2012).

[73] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **118**, 2309 (1996).

[74] M. Cossi, N. Rega, G. Scalmani, and V. Barone, J. Comput. Chem. **24**, 669 (2003).

[75] H. Li, C. S. Pomelli, and J. H. Jensen, Theor. Chem. Acc. **109**, 71 (2003).

[76] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H. Van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, 2nd ed. (SIAM, Philadelphia, PA, 1994).

[77] V. Barone, M. Cossi, and J. Tomasi, J. Chem. Phys. **107**, 3210 (1997).

[78] J. L. Pascualahuir, E. Silla, and I. Tunon, J. Comput. Chem. **15**, 1127 (1994).

[79] M. J. T. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, Gaussian 03, Revision E.01, Gaussian, Inc., Wallingford, CT, 2004.

[80] R. W. Pastor, B. R. Brooks, and A. Szabo, Mol. Phys. **65**, 1409 (1988).

[81] H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).

[82] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).

[83] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, J. Comput. Chem. **26**, 1668 (2005).

[84] E. Lin and M. S. Shell, J. Chem. Theory Comput. **5**, 2062 (2009).

[85] L. L. Duan, Y. Gao, Y. Mei, Q. G. Zhang, B. Tang, and J. Z. H. Zhang, J. Phys. Chem. B **116**, 3430 (2012).

[86] R. B. Best and G. Hummer, J. Phys. Chem. B **113**, 9004 (2009).

[87] C. Curutchet, M. Orozco, and F. J. Luque, J. Comput. Chem. **22**, 1180 (2001).

[88] See supplementary material at http://dx.doi.org/10.1063/1.4833678 for Table S1 that shows the deviation of the calculated total electrostatic solvation energy between the EE-GMFCC-CPCM method using PPC charges and the standard full system calculations at the HF/6-31G* level.

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 128.122.149.145 On: Fri, 17 Jan 2014 07:57:10