JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

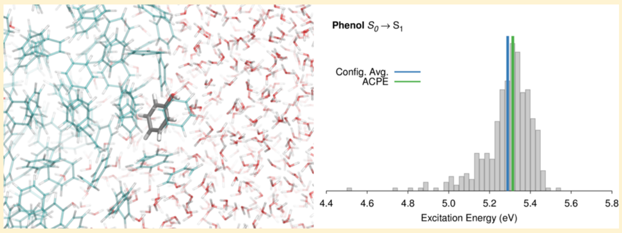
Averaged Condensed Phase Model for Simulating Molecules in Complex Environments

Dominique Nocito and Gregory J. O. Beran*

Department of Chemistry, University of California, Riverside, California 92521, United States

S *Supporting Information*

ABSTRACT: The need for configurational sampling dramatically increases the cost of combined quantum mechanics/molecular mechanics (QM/MM) simulations of chemical processes in solution. We developed an averaged condensed phase environment (ACPE) model that constructs an effective polarizable environment directly from explicitly sampled molecular dynamics configurations via the K-means++ algorithm and a mathematically rigorous translation of the molecular mechanics parameters. The model captures detailed heterogeneous features in the environment that may be difficult to describe using a conventional polarizable continuum model. Instead of performing repeated QM/MM calculations for each new configuration of the environment, the ACPE approach allows one to perform a single QM calculation on an averaged configuration. Here, we demonstrate the model by computing electronic excitation energies for several small molecules in solution. The ACPE model predicts the excitation energies in excellent agreement with conventional configurational averaging yet with orders of magnitude of reduction in the computational cost.



1. INTRODUCTION

Combined quantum mechanical/molecular mechanics (QM/MM) calculations provide an effective route toward modeling complex systems in the condensed phase with much less computational effort than fully QM simulations. Nevertheless, modeling condensed phase systems such as a molecule in solution remains challenging due to the need to perform the QM/MM calculations on large numbers of sampled configurations. QM/MM configuration sampling can be performed directly with QM/MM dynamics or indirectly by first sampling with MM and subsequently performing QM/MM calculations on configurations extracted from the MM ensemble. Even in the latter approach, repeated QM calculations are needed for each change in the MM environment. Strategies that reduce the computational effort associated with evaluating the response of the QM region to the environment are therefore important.

One might circumvent the need to perform a new QM calculation for each new configuration by either approximating or precomputing the response of the QM solute to the solvent using point-charge or more elaborate representations of the solute.[1−9] More recently, Sodt et al. proposed a family of multiple environment, single system (MESS) models which use an efficient correction to update the QM energy and orbitals/density in response to a change in the MM environment.[10] For example, the Hessian (H)-based MESS-H variant estimates the QM/MM energy in a new solvent configuration from the energy of a previous configuration based on a single Newton−Raphson step update of the Kohn−Sham orbitals. The approximate orbital Hessian used in the Newton−Raphson step needs to be computed only once and can be reused for each new configuration of the environment.

Polarizable continuum models (PCMs) lie at the opposite extreme. Rather than explicitly sampling configurations of the environment, PCMs represent the environment as a bulk dielectric medium.[11−13] Polarizable continuum models often do an excellent job of capturing bulk solution behaviors, but they perform more poorly when specific, local solute−solvent interactions are important. For example, the Cope elimination reaction rate accelerates a million-fold upon switching from a protic to aprotic solvents.[14−16] Hydrogen bonding between the solute and protic solvent molecules preferentially stabilizes the reactant, effectively increasing the activation barrier and slowing the reaction rate. Inclusion of key explicit solvent molecules inside the solute cavity is necessary to capture this effect with a PCM model.[17] Continuum models also have difficulty describing inhomogeneous environments for which a bulk dielectric is ill-defined. The effective dielectric constant of a protein has been frequently debated, for instance.[18−21]

An interesting family of methods lies in between explicit QM/MM evaluation of sampled configurations and polarizable continuum models. In these methods, one embeds the QM calculation in an averaged or effective representation of the environment. As in a PCM approach, replacing hundreds or more QM/MM calculations with a single calculation in an averaged environment reaps massive computational savings. At the same time, constructing the averaged environment from explicit configurations can retain essential features that might otherwise be lost in a bulk continuum approximation. These methods do assume that the response of a system to its

Received: September 8, 2016
Published: February 7, 2017

DOI: 10.1021/acs.jctc.6b00890

averaged environment is consistent with taking the average over many individual responses of the system to different instantaneous environments. Though there may be situations where this approximation does not behave well, it often appears to be a useful one.

One such model, the averaged solvent electrostatic potential (ASEP) model developed by Aguilar and co-workers, embeds a solute monomer in a field of point charges fitted to reproduce the average electrostatic potential felt on the solute due to the environment.[22−25] Another model, the three-dimensional reference site interaction model (3D-RISM) approach,[26,27] allows the computationally efficient evaluation of solvent density distributions and thermodynamic parameters without requiring explicit solvent simulations. RISM approaches can be combined with QM simulations to study solvation effects.[27−29]

Previously, we presented a mean-field model that employs a mathematically rigorous coarse graining of the environment.[30] This coarse-grained (CG) model constructed a radial grid of CG points about the solute and then averaged the effective force field parameters at these grid points over space and time. The coarse graining relied on formally exact spherical harmonic translation formula to translate the MM parameters from their explicit atomic sites to these CG grid points. These translations are analogous to the ones used in the fast multipole method,[31] for example, except in this case, they were applied to multipoles (electrostatics), polarizabilities (induction), and frequency-dependent polarizabilities (van der Waals dispersion). The resulting translated MM parameters at each CG grid point are summed and then averaged over the ensemble of configurations. The use of a grid of effective polarizable multipoles to represent the solvent is also akin to the Langevin dipole solvation model.[32−36] In the Langevin dipole model, the magnitude and orientation of the dipole at each grid point are optimized simultaneously with the wave function of the solute. Whereas the Langevin dipoles model has primarily been parametrized for aqueous solution, our CG approach can be applied to an arbitrary molecular environment more readily.

Here, we extend that earlier CG model in two key ways. First, we improve the manner in which the coarse-grained points are chosen. The previous grid approach proved too sensitive to the specific grid-point locations. Instead of enforcing a regular grid, the current work employs clustering algorithms to place CG site locations "naturally" based on the explicit locations of atoms/molecules in the sampled configurations. Second, we improved the physical behavior of the coarse-grained polarization model. In particular, to retain distinctions in the polarizabilities of different atoms or molecules, the coarse-graining is now performed separately over each unique atom type. Furthermore, the atomistic model employed here would typically include only intermolecular polarization because intramolecular polarization is already accounted for in the multipolar expansion. However, the earlier CG approach lost the distinction between inter- and intramolecular polarization. To regain some of that distinction in the coarse-grained representation, nearby CG sites are now clustered, and polarization occurs only between clusters rather than within them. These clusters loosely correspond to the dynamic region inhabited by a given solvent molecule during the simulation.

We examine the performance of the refined averaged condensed phase environment (ACPE) model by computing excitation energies of small organic molecules in solution. We demonstrate that the ACPE model maintains important features of the underlying solvent structure and that it can be used to describe inhomogeneous features in complex environments that would be difficult to describe with a conventional, homogeneous polarizable continuum model. At the same time, the predicted ACPE excitation energies in the averaged environment agree very well with those from a more traditional QM/MM average over many configurational snapshots. Importantly, the predicted excitation energies prove fairly robust to variations in the specific CG sites generated by the ACPE model.

2. THEORY

The ACPE procedure consists of six main steps, as illustrated by the flowchart in Figure 1:

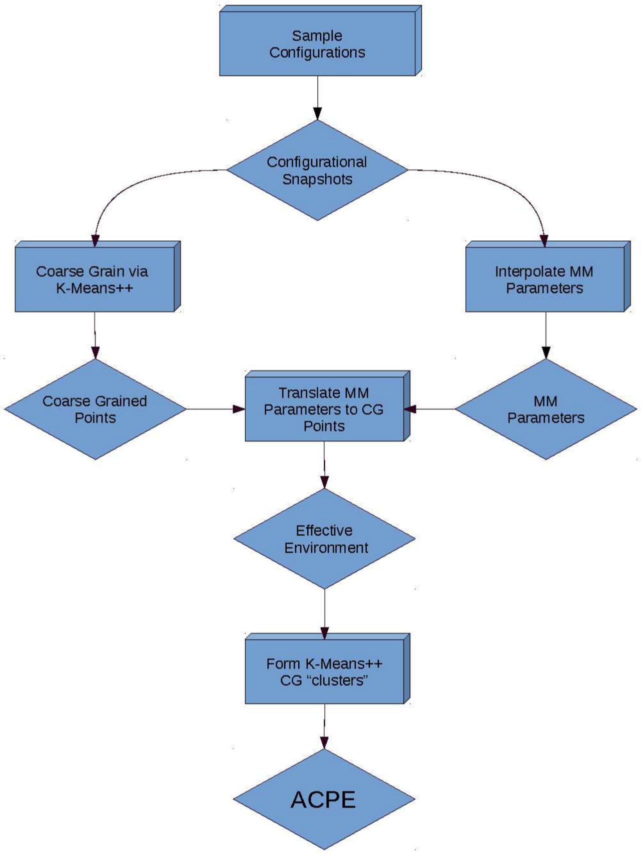

Sample Configurations
Configurational Snapshots
Coarse Grain via K-Means++
Interpolate MM Parameters
Coarse Grained Points
Translate MM Parameters to CG Points
MM Parameters
Effective Environment
Form K-Means++ CG "clusters"
ACPE

Figure 1. Flow-chart outlining the steps involved in the ACPE model.

1. Sample the configurational space of the system.
2. Superimpose the atomic coordinates of sampled configurations.
3. Generate a set of CG sites via the K-means++ algorithm.
4. Obtain MM parameters for each molecule via interpolation.
5. Translate atomic MM parameters to the nearest CG site of the same atom type.
6. Group CG sites into CG clusters via a second round of K-means++.

Step 1 uses standard molecular dynamics, Monte Carlo, or related techniques to sample the configurations of the environment. In all cases here, the molecule of interest (the "solute") is frozen at a fixed geometry during the configurational sampling to allow straightforward superposition of the sampled environment ("solvent") configurations in step 2. This approximation has been used in the MESS[10] and ASEP models[37] as well. In principle, one could repeat steps 1−6 for various solute geometries if solute dynamics are also important.[37,38] Step 2 involves merging the atomic coordinate files for the molecules in the environment over all sampled configurations into a single list. Step 4 assumes that the force field parameters can vary as a function of the specific molecular geometries in the environment. Here, we vary the water force field parameters with intramolecular geometry. If the MM parameters are constant, step 4 can be skipped. Steps 3−6 are described in more detail below.

2.1. Determination of Coarse Graining Sites.

Step 3 in the ACPE procedure automatically determines the locations of the CG sites in the environment region based on the atomic coordinates in sampled configurations via K-means clustering. K-means is widely used in machine learning, pattern recognition, and data-mining.[39] In general, K-means clusters a data set of n points into k groupings. Here, we adapt K-means to automatically cluster n atoms into k coarse-grained sites.

Distinct sets of CG sites are determined here for each symmetrically unique atom type present in the environment. For an aqueous environment, for instance, K-means coarse-graining is applied separately for the solvent oxygen and hydrogen atoms. This preserves the physical behaviors of different atom types and retains aspects of molecular structure within the CG model. The number of CG sites k is chosen as the average number of atoms per configuration of the type being coarse-grained times a user-selected scaling factor. As determined by empirical testing, scaling factors of 1−10 for light atoms and 5−40 for heavier atoms work well in the examples considered here, as will be discussed further in Section 4.1. Further study is needed to develop a more universal algorithm for choosing the scaling factor. Although this coarse-graining utilizes many more sites than are present in a single configuration of the initial system, it contains orders of magnitude fewer sites than the total number of sites found across the hundreds (or more) configurations being averaged over. In other words, this approach coarse-grains over both spatial coordinates and configurational snapshots simultaneously, in contrast to a more traditional coarse-graining that is purely spatial.

The K-means algorithm seeks to identify the set of CG sites which minimizes the sum of the distances between the atomic sites and their nearest CG site. Specifically, it assigns each atom in the atomistic picture to a CG site and seeks to minimize the objective function:

$$\underset{S}{\arg\min} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{1}$$

where S is the set of all atoms, x is the position of a single atom in S, S_i is the set of atoms associated with the ith CG site, and μ_i is the centroid (position) of the ith CG site. Finding the globally optimal solution to the problem of clustering n points into k groups scales as $O(n^{(d+2)k+1})$, where d is the dimensionality of each point (three for Cartesian coordinates).[40] The K-means algorithm iteratively searches for the solution to this problem in $O(nkdi)$ effort, where i is the number of iterations needed to converge. Inclusion of a neighbors list (described below) can be used to effectively eliminate k from the scaling. Because the K-means algorithm used here relies on Euclidean distance between the atoms and CG sites, the individual atoms clustered into a single CG site tend to be roughly spherical.

In this paper, we employ a variant of the K-means algorithm known as K-means++, which differs from K-means only in the initialization step. In traditional K-means, a poor initialization of the CG sites can lead to poor clustering. K-means++ addresses this problem by biasing the initialization procedure toward evenly distributed CG sites.[41] The K-means++ initialization typically leads to faster convergence of the algorithm and more optimal solutions.[41] The K-means++ initialization step is performed as follows:

1. Select one atom uniformly at random to be a CG site.
2. Compute $d(A)$, the square of the distance between each atom and its nearest centroid.
3. Weight the probability that each remaining atom A will be chosen as the next CG site by

$$\frac{d(A)}{\sum_{A=1}^{n} d(A)}$$

4. Choose the next CG site at random and repeat steps 2−4 until k CG sites have been initialized.

Step 3 increases the probability that the initial guess CG sites will be well-separated throughout the environment. Note that while this K-means++ guess initializes CG sites on individual atoms, the final CG sites obtained upon converging the K-means clustering algorithm are unconstrained and can lie anywhere in space.

A neighbors list was implemented to reduce the number of distance calculations required during the K-means clustering process. The neighbors list defines a sphere of inclusion around every atom, and only CG sites that lie within this sphere are considered in the clustering of that atom. This is similar to a Verlet neighbors list[42] used in molecular mechanics calculations to keep a list of all neighboring particles for which interactions will be calculated. Assuming uniformly distributed CG sites, the neighbors list reduces the number of distance calculations per atom from k to $\rho\pi r^3$, where r is the radius of the neighbors list and ρ is the density of centroids. For the examples considered in this paper, using the neighbors list accelerates the K-means algorithm by an order of magnitude, but this improvement is ultimately dependent on k and the size of the system. With the neighbors list included, the K-means algorithm runs as follows:

1. Determine neighbors list of CG sites lying within 3 Å of each atom.
2. Compute distance squared $\|x - \mu_i\|^2$ between each atom x and the CG sites μ_i associated with x in the neighbors list.
3. Assign each atom to its nearest CG site.
4. Compute new CG site positions as the mean of the positions of atoms assigned to it.
5. Calculate the sum of the absolute change in positions of the CG sites from their previous position.
6. Recompute the neighbors list if the sum from step 5 has reached a defined threshold.
7. Repeat steps 2−6 until the sum from step 5 equals zero.

For step 6, the neighbors list is updated if the average change in position of the CG sites exceeds half the radius of the sphere in step 1 (i.e., 1.5 Å). This K-means clustering is applied to the complete set of superimposed atomic configurations.

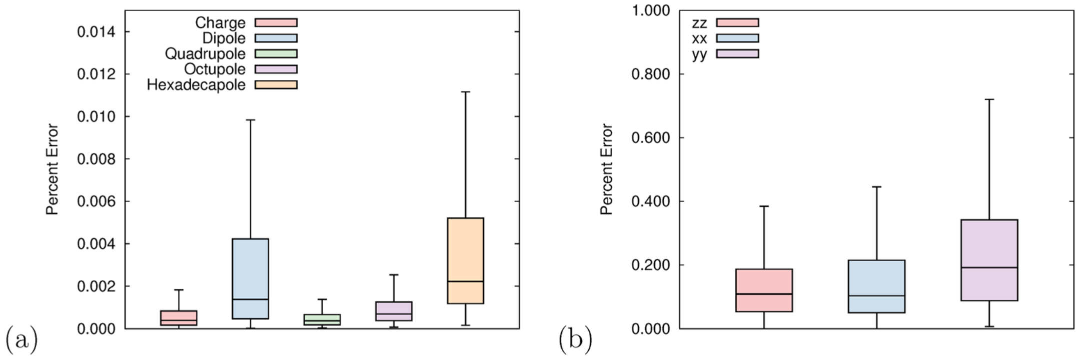


Figure 2. Box-plot distributions of absolute percent errors in the (a) multipole moments and (b) polarizabilities between directly computed and the interpolated AIFF parameters for 400 water geometries. The boxes indicate the median error (center line) and the central 50% of the data, while the whiskers indicate the largest errors.

2.2. Force Field Parameter Interpolation.

Once the set of k CG sites has been determined via K-means++, parameters are needed to describe the interactions between the system and its coarse-grained environment. That requires obtaining force field parameters for the species in the original atomistic representation of the environment (e.g., for each solvent molecule in each sampled configuration) and then mapping those parameters onto the coarse-grained representation (Section 2.3).

The success of any embedding treatment is dependent on the manner in which the environment is modeled. The molecules in the environment here are modeled using a polarizable ab initio force field (AIFF) which has been demonstrated to perform well for describing long-range and many-body interactions.[43−46] The AIFF is parametrized in terms of atom-centered distributed multipoles (electrostatics),[47−49] distributed polarizabilities (polarization),[50−52] and distributed frequency-dependent polarizabilities (van der Waals dispersion).[53] In the examples studied here, we perform electrostatic embedding only, so the dispersion contributions are ignored. However, dispersion contributions were included in our earlier coarse-graining work.[30]

The AIFF parameters are typically calculated on the fly from density functional theory (DFT) for each molecule in its current geometry. Using CamCASP,[54] calculating these parameters for a water molecule typically takes a few minutes. Performing such calculations over hundreds of solvent molecules in hundreds of configurations quickly becomes computationally demanding. Instead, we precomputed the water parameters at 20 water bond angles and 20 bond lengths for each O−H bond (i.e., 8000 geometries total). The force field parameters (distributed multipoles and polarizabilities) can then be interpolated from this grid of configurations using a "natural" cubic spline.

The spherical tensor multipole and polarizability force field parameter data on the interpolation grid is stored in a local coordinate frame. Mapping the interpolated force field parameters onto each individual molecule in the environment requires rotating the local-frame parameters into the global coordinate system. Rotations of the multipole moments and polarizabilities are performed using explicit expressions.[55] The rotation matrix elements are expressed as polynomials of degree $\leq l$, where l is the rank of moment being rotated in terms of the elements of the 3 × 3 rotation matrix. Rotation matrixes for up to l = 4 (hexadecapole) are tabulated in the Supporting Information.

Overall, interpolation of the force field parameters provides excellent accuracy at a tiny fraction of the computational cost of computing the parameters directly. Figure 2 presents the errors in the multipole moments and polarizabilities arising from the interpolation for 400 solvent water conformations taken from an MD simulation. Errors in the multipoles are typically well below 0.01%, while errors in the polarizabilities are no more than a few tenths of a percent. At the same time, this interpolation procedure reduces the computational cost of obtaining AIFF parameters dramatically. In a test job containing 1000 solvent configurations of 1600 water molecules (i.e., 1.6 million waters total), interpolation lowers the cost of generating the force field parameters from 9.1 years (at ~3 min each) to only 9.7 h of CPU time. Each interpolation is independent of the others, so they can be performed in a highly parallel fashion if desired.

2.3. Force Field Parameter Translation.

Once the set of CG sites has been determined via the K-means++ algorithm (Section 2.1) and force field parameters have been obtained for each molecule in the original explicit representation of the environment (Section 2.2), the force field parameters for each atom are translated to its associated CG site. Parameters are summed at each CG site and divided by the number of configurations N to obtain average values. As described previously,[30] the translation exploits the fact that a multipole moment $Q_{l'k'}$ at a new point in space C can be exactly represented as a linear combination of multipoles Q_{lk} at the original point O.[56] The functional form for the translation of the moments is

$$Q^{C}_{lk} = \sum_{l'=0}^{l} \sum_{k'=-l'}^{l'} \left[\binom{l+k}{l'+k'} (l - kl' - k') \right]^{1/2} Q^{O}_{l'k'} R_{l-l',k-k'}(-c) \tag{2}$$

where the Q^{C}_{lk} are the multipole moments at the final position, the terms in curved brackets are binomial coefficients, $Q^{O}_{l'\,k'}$ are the multipole moments at the initial location, and $R_{l-l',k-k'}(-c)$ is a regular spherical harmonic. If k is not equal to zero, the resulting multipole moment will be complex. Real multipole moments can be constructed according to

$$R_{lm} = \frac{(R_{lmc} + iR_{lms})}{2b_m} \tag{3}$$

where b_m is a piece-wise defined coefficient, R_{lmc} and R_{lms} are the regular spherical harmonics. Additional details for deriving the translation expressions and a complete set of translations for up to hexadecapole moments are listed in the Supporting Information.

In this approach, multipolar translations are expressed as polynomials of degree l in terms of the elements of the translation vector with coefficients of the moments of rank $\leq l$. A charge distribution described by a finite number of moments at a point would require moments up to infinite rank to completely describe it at another point. For computational expediency, we truncate the multipole expansion at hexadecapoles ($l = 4$). In principle, errors introduced by translating the multipole moments to the CG sites could be systematically reduced by including higher-order moments.

The translation of the polarizabilities can be determined by applying eq 2 to the multipolar operators that occur in the formula for the polarizability.[30] For example, for the dipole−dipole polarizability tensor elements α_{tu} are given by,

$$\alpha_{tu} = \sum_{n}{}' \frac{\langle 0|\hat{\mu}_t|n\rangle\langle n|\hat{\mu}_u|0\rangle + \langle 0|\hat{\mu}_u|n\rangle\langle n|\hat{\mu}_t|0\rangle}{W_n - W_0} \tag{4}$$

where $\hat{\mu}_u$ and $\hat{\mu}_t$ are different components of the dipole moment operator and $|0\rangle$ and $|n\rangle$ are the ground and excited states of the system, respectively, with energies W_n. For example, the tth component of the dipole moment operator translated from initial point O to some new point C is given by

$$\hat{\mu}_t^C = \hat{\mu}_t^O + qC \tag{5}$$

where $\hat{\mu}_t^C$ is the tth element of the translated dipole operator, $\hat{\mu}_u^O$ is the original tth element of the dipole operator, q is the charge of this site, and C is the tth element of the translation vector. Substituting the operator form for this expression in for the dipole operators in eq 4, one finds that because qC term is constant and the eigenstates are orthogonal, matrix elements involving the charge q are zero by orthogonality. In other words, the translated dipole−dipole polarizability is invariant to translation. See the Supporting Information of ref 30 for details. Note that polarizability tensor elements involving higher-rank contributions are not invariant; however, only dipole−dipole polarizabilities are used in the embedding model here. Nevertheless, polarizability translation expressions up to rank 2 (quadrupole−quadrupole) are provided as Supporting Information.

2.4. Clustering of Coarse-Grain Sites.

The K-means++ coarse graining in Section 2.1 produces a dense grid of points that can accurately reproduce the electrostatic interactions between the solute and environment. Figures 3a and b compare an individual water solvent configuration and the cluster of oxygen and hydrogen coarse-graining points representing an ACPE constructed by averaging over 400 solvent configurations. However, the coarse graining eliminates the definitions of individual molecules in the environment, which blurs the distinction between intra- and intermolecular polarization. Intramolecular polarization is already implicitly included in the AIFF monomer distributed multipoles, while intermolecular polarization needs to be modeled explicitly.

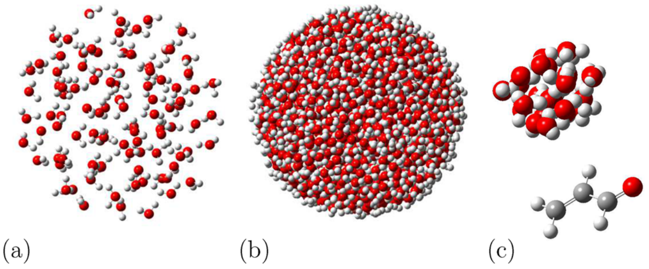


Figure 3. (a) Single configuration snapshot of the water solvent. (b) ACPE environment from 400 sampled configurations. (c) A single ACPE molecular cluster near an acrolein solute.

To recapture some of the distinction between intra- and intermolecular and enable intermolecular polarization in the coarse-grained representation, a second round of K-means++ is employed to group CG sites into CG clusters. A given CG cluster is composed only of CG sites derived from atoms from a given type of molecule. For example, in the mixed water/benzene environment described in Section 4.4, a given CG cluster would involve only oxygen and hydrogen sites derived from water molecules or carbon and hydrogen sites derived from benzene molecules. A given CG cluster loosely corresponds to the dynamic domain sampled by a single molecule (though it may be derived from contributions from various molecules). The ACPE model treats polarization only between CG clusters. Polarization within a CG cluster is forbidden.

Figure 3c shows a single ACPE water CG cluster interacting with an acrolein solute molecule. Note that while the CG sites in the CG clusters may seemingly resemble water molecules, the actual bond distances and angles between oxygen and hydrogen sites correlate only loosely with real water molecules. It is also worth emphasizing that these water-like distributions of hydrogen and oxygen CG sites arise “naturally” from the K-means++ algorithm. The model was not steered to produce water-like CG sites.

Two parameters are used to define the CG clusters and their interactions. The first parameter is the number of CG clusters, K. We choose K to equal the average number of the solvent molecules from which the cluster was derived. For instance, if a solute is surrounded by 256 water molecules in each configuration being averaged over, there will be 256 CG clusters in the final ACPE.

The second parameter is the minimal distance between points in different clusters for which polarization is allowed. The goal is to allow maximal polarization while avoiding the polarization catastrophe. To determine this, ACPE calculates self-consistent atom-centered induced dipoles due to many-body polarization. ACPE then calculates the average induced dipole for each atom type. For atoms with induced dipoles less than the system average, the cutoff is decreased by 0.1 Å. For atoms with induced dipoles greater than 0.03 au, it is increased by 0.1 Å. This process of calculating the polarization and examining the induced dipoles is repeated until the many-body induction energy is between 95−105% of the configurational average of the original atomistic model or until 10 iterations have been reached. These calculations are performed purely at the MM level, so they can be done inexpensively.

This procedure was derived empirically, but it ensures that the coarse-grained polarization model faithfully reproduces the original atomistic polarization while avoiding the polarization catastrophe from close-lying clusters. As shown in Figure 4, the majority of CG sites are located on the interior of CG clusters and are sufficiently far from atoms in other CG clusters that they have a polarization cutoff of 0.0 Å (i.e., full polarization). Nonzero polarization cutoffs are primarily needed for atoms on the edges of adjacent CG clusters. In the current implementation, the polarization cutoff is a hard step function:

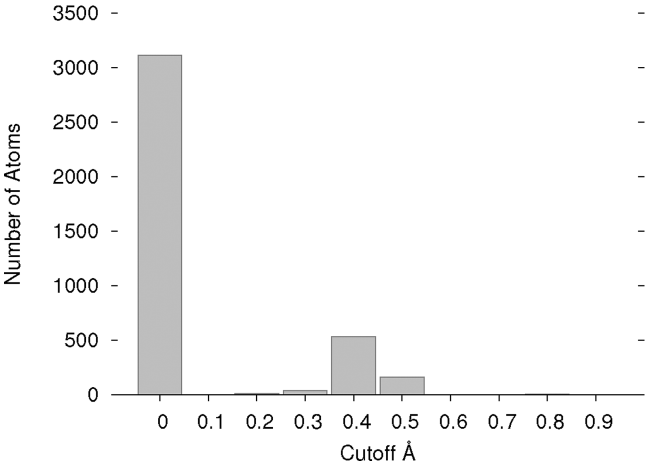


Figure 4. Distribution of polarization cutoff distances for the aqueous environment around acrolein.

polarization is either allowed or not. One might instead use a smooth damping function to interpolate between complete polarization and no polarization, but that is not investigated here.

In principle, the ideas here can be applied to any fixed-charge, multipolar, or polarizable force field which is derived from a multipolar expansion. However, because translation of the force field parameters introduces higher-order multipolar components, it complicates the force field model. Translating atomic parameters from a simple fixed-charge force model like all-atom OPLS, for instance, leads to the introduction of dipoles and higher-order terms into the electrostatic model. Similarly, translating polarizable force field parameters from the Amoeba model[57,58] introduces higher-rank multipoles and polarizabilities. In practice, one truncates the multipolar expansions to ignore some of these higher-rank contributions, albeit with some loss in accuracy. In the case considered here, we maintain the original force field expansions with maximal rank 4 (hexadecapole) multipoles and rank 1 (dipole−dipole polarizabilities) in the coarse-grained model. This captures the leading contributions arising from the translated dipoles, quadrupoles, and octupoles but neglects higher-rank contributions from translating hexadecapoles. Similarly, we neglect any changes in the polarizability beyond dipole−dipole that arise from translation. In other words, the ideas here are best-suited for force fields that already include multipoles beyond point charges.

2.5. Molecular Excitation Energies with Polarizable Embedding. In the end, the ACPE procedure described above produces a set of multipole moments and polarizabilities which can then be used to construct an embedding environment for a variety of quantum mechanical calculations. For the purposes of this paper, the ACPE model is used to construct a polarizable solvent environment for computing solute excitation energies with time-dependent density functional theory (TDDFT).

All embedding calculations use a self-consistent polarizable embedding (PE) scheme,[59] which is implemented in Dalton 2013.[60] This model allows polarizable embedding with point multipoles and polarizabilities. The electrostatic potential is modeled with multipole moment up to rank 4 (hexadecapole), which is sufficient to model the permanent charge distribution. Polarization is treated using anisotropic dipole−dipole polarizabilities.

3. COMPUTATIONAL METHODS

3.1. Configuration Sampling. Configuration sampling was performed via molecular dynamics (MD) simulations using Tinker 7.1[61] and the OPLS-AA force field.[62] In the first three test systems (s-trans acrolein, acetone, and pyrimidine) in aqueous solvent, MD simulations were performed under periodic boundary conditions in a cell containing a single solute molecule and 1600 water molecules. Five-hundred picoseconds of NPT dynamics at 298.15 K and 1.0 atm were carried out to equilibrate the system, followed by a 1.0 ns NVT production run. The solute molecule was held fixed during the MD simulations at a geometry optimized using the CAM-B3LYP functional, aug-cc-pVTZ basis, and implicit water solvation (using the integral equation formalism polarizable continuum model[63] in Gaussian 09).[64] Time steps of 1.0 were used throughout. Four-hundred configurations were sampled at intervals of 1.0 ps over the last 0.4 ns of the production run.

For the benzene/water interface example, a rectangular box consisting of a single, frozen phenol molecule at the interface of 1600 waters and 138 benzene molecules stacked along the z-coordinate was generated (see Figure 8). Periodic boundary conditions were employed. The system was allowed to equilibrate for 100 ps under NPT conditions. The phenol was then frozen, and 400 ps of additional NPT equilibration were carried out to obtain a box with lengths 36.02, 36.02, and 55.31 Å along x, y, and z, respectively. Subsequently, 1.0 ns of NVT MD production run was carried out. Again, 400 configurations were sampled at 1.0 ps intervals over the last 0.4 ns of the NVT simulation.

Finally, a spherical solvation shell consisting of all solvent molecules lying within 9 Å of any atom in the solute molecule was extracted from each MD configuration. These large clusters were then used to construct the ACPE model. These finite clusters may not fully capture bulk solvation effects, but they provide a useful test for the ACPE coarse-graining procedure. One could potentially employ larger clusters or in some cases surround them with a bulk continuum model for longer-range effects, though we do not do so here. As shown in Table S1 of the Supporting Information, increasing the cluster radius to 15 Å alters the excitation energies in a single acrolein configuration by 0.1 eV or less.

3.2. ACPE Construction. The ACPE for a given system was constructed from the 400 configurations sampled from the MD. Unless otherwise mentioned, scaling factors of 5 (H atoms) and 30 (heavy atoms) were used to define the number of CG sites k used in the initial K-means++ coarse-graining for each system. This means, for example, that in a system with an average of 100 water molecules (100 oxygen and 200 hydrogen atoms) per configuration, there would be 5 × 200 = 1000 H sites and 30 × 100 = 3000 oxygen coarse-graining sites. These scaling factors were chosen empirically based on a survey of scaling parameters for acrolein (see Section 4.1 for details).

The AIFF distributed multipoles and polarizabilities for water and benzene were computed with CamCASP version 5.6[54] using asymptotically corrected PBE0 and the Sadlej basis. An ionization potential of 0.4638 au was used for the PBE0 asymptotic correction of water. For water, the force field parameters at each geometry were interpolated as described in Section 2.2. For benzene, force field parameters computed at the equilibrium geometry were used throughout. Multipole moments up to hexadecapoles (rank 4) on heavy atoms and dipoles (rank 1) on hydrogen were used along with

Table 1. Predicted B3LYP/aug-cc-pVDZ ACPE Excitation Energies for Acrolein with Different Scale Factors for the Number of Heavy and Light Atom CG Sites[a]

		$(n \to \pi^*)$ light						$(\pi \to \pi^*)$ light			
		5	SD	10	SD			5	SD	10	SD
heavy	5	3.79	0.01	3.79	0.00	heavy	5	5.95	0.00	5.96	0.01
	10	3.82	0.02	3.80	0.03		10	5.94	0.00	5.94	0.03
	15	3.82	0.01	3.75	0.02		15	5.92	0.02	5.98	0.01
	20	3.88	0.02	3.85	0.03		20	5.90	0.01	5.92	0.01
	25	3.83	0.03	3.91	0.04		25	5.94	0.01	5.90	0.03
	30	3.85	0.02	3.91	0.06		30	5.91	0.02	5.89	0.03
	35	3.80	0.03	3.90	0.08		35	5.92	0.01	5.87	0.04
	40	3.82	0.05	3.88	0.04		40	5.94	0.02	5.91	0.02

[a] Each ACPE calculation was repeated four times with different random initialization. Values report the average excitation energies and standard deviations. For reference, averaging over the 400 configurations explicitly predicts excitation energies of 3.84 eV ($n \to \pi^*$) and 5.86 eV ($\pi \to \pi^*$).

polarizabilities up to dipole−dipole (rank 1) The translated multipoles and polarizabilities were truncated at ranks 4 and 1, respectively.

3.3. Excitation Energy Calculation. TDDFT excitation energies were computed using the polarizable embedding (PE) module in Dalton 2013[60] using the CAM-B3LYP density functional[65] and the aug-cc-pVTZ basis.[66] For the CAM-B3LYP functional, the parameters α modify the fraction of HF exchange, and β modifies the fraction of the DFT exchange for short- and long-range interactions. Here, α and β values of 0.19 and 0.46 were used, respectively.[67] The switching factor between HF and DFT exchange μ is equal to 0.33, as proposed in the original work.[65] For the purposes of exploring the effect of the ACPE parameters in Section 4.1, calculations were performed with the less-expensive B3LYP/aug-cc-pVDZ model.

The PE was modeled with multipole moments up to hexadecapole (rank 4) and anisotropic dipole−dipole polarizabilities (rank 1). Polarization was treated self-consistently among the ACPE CG clusters and the QM region. Polarization within a CG cluster is omitted, and short-range polarization cutoffs between CG sites in different clusters were implemented as described in Section 2.4.

3.4. ACPE Validation. For comparison purposes, distributions of excitation energies were also computed using the polarizable embedding model for each of the 400 individual explicit configurations for each system. A “configurational average” excitation energy is obtained by computing the mean excitation energy for each state over the 400 configurations. In some examples, integral equation formalism PCM calculations using default parameters for an aqueous environment were also performed with Dalton. Including a small cluster of explicit solvent molecules might improve the PCM results, but this was not done here to ensure that the QM regions are identical in all calculations performed here.

4. RESULTS AND DISCUSSION

To investigate the performance of the ACPE, we first examine the structure of the averaged environment generated by the model. Next, we compute low-lying vertical singlet excitation energies for several small molecules in aqueous solution. Finally, to demonstrate application of the ACPE model to a more complicated, inhomogeneous environment, we study the excitations of a phenol molecule residing at the interface between benzene and water solvents. The sharp differences in solvent polarity and the spatial phase separation exhibited by the two solvents would make this type of system much harder to describe with conventional implicit solvent models. Even if a few of the closest solvent molecules were modeled explicitly, it is not obvious how to handle the two very different bulk dielectrics simultaneously in a traditional PCM model.

4.1. Determination of the ACPE Model Parameters. The ACPE coarse-graining procedure contains a number of potential parameters that might affect the model results. First, the initial CG sites in the K-means++ algorithm are determined randomly, which raises the question of the reproducibility of the results for different random seeds. Second, one must choose the density of CG sites, which is defined as some scaling factor times the average number of atoms of a given type in the MD configuration snapshots. Third, short-range damping is used to avoid the polarization catastrophe in the embedding procedure, as described in Section 2.4.

To explore how the first two parameters affect the excitation energies in the ACPE environment, we consider the lowest two excited states of acrolein in water. As a reference, we first computed the excitation energy with B3LYP/aug-cc-pVDZ via polarizable embedding for each of 400 solvent configurations. Each configuration consists of a single acrolein surrounded by an average of 97 water molecules. Averaging over the 400 configurations, we obtain average excitation energies of 3.84 eV ($n \to \pi^*$) and 5.87 eV ($\pi \to \pi^*$).

Next, we performed ACPE calculations with four different random seeds (i.e., distinct K-means++ initializations) and varying scale factors for the number of heavy (oxygen) and light (hydrogen) atom CG sites (Table 1). For any given set of light/heavy atom scale factors, the standard deviation in the predicted excitation energies due to different random seeds in the K-means++ initialization is less than 0.1 eV.

Similarly, the excitation energies are relatively insensitive to the heavy and light atom scale factors. Using larger scale factors (more CG sites) reduces the typical distance between the original atom and its assigned CG site. Translating the force field parameters (multipoles and polarizabilities) to shorter distances reduces the magnitude of the higher-rank components introduced upon translation. The embedding model here includes up to hexadecapoles for the permanent multipole moments. Thus, the higher-order components introduced by translation are described fairly well. However, the PE model supports only dipole−dipole polarizabilities, so important higher-rank contributions to polarization introduced by longer translation distances will be omitted.

Higher-rank distributed multipoles can be significant in magnitude on heavy atoms,[68] which in turn means it may be beneficial to translate their parameters less distance (to minimize the introduction of components with rank >4 that are not included in our model). For hydrogen, the force field representation before translation includes only up to dipoles (rank 1),[43,44] so the higher-order contributions introduced by translation are captured more completely by the final rank 4 representation of the embedding environment. Despite these considerations, there does not appear to be any clear preference for certain combinations of scale factors in practice, as seen in Table 1. Scale factors of 30 for heavy atoms and 5 for light atoms seem to behave well and are used for all calculations described below.

With these parameters, the mean distance between the original atoms and their corresponding CG sites in the four ACPE runs described above is 0.42 ± 0.14 Å for oxygen and 0.69 ± 0.22 Å for hydrogen, respectively. Those individual CG sites are grouped into CG clusters (Section 2.4). If both the explicit solvent molecules and the CG sites were uniformly distributed, each heavy or light atom CG cluster would contain 30 (oxygen) or 2 × 5 (hydrogen) CG sites (i.e., the number of sites would match the scale factors). In the four ACPE runs described above, the clusters averaged 30.0 ± 3.9 CG sites for oxygen and 10.0 ± 1.6 for hydrogen.

Our implementation of the ACPE algorithm has not been fully optimized for computational efficiency. Nevertheless, timings of the current implementation demonstrate that the construction of the ACPE requires only a fraction of the subsequent excitation energy calculation. Table 2 breaks the timings down into individual components of the ACPE algorithm. The large number of polarizable sites employed in the polarizable embedding with ACPE also modestly increases the time for the TDDFT calculation (by ∼3000 s). Overall, generating the ACPE and computing the ten lowest excited states of acrolein in the water via the ACPE model requires 26 000 s compared to 16 000 s for embedding with multipoles and polarizabilities from a single configuration (i.e., without any ACPE model). In other words, employing the ACPE model allows one to mimic the effect of hundreds of solvent configurations at a cost only 60% higher than that of a TDDFT calculation on a single configuration snapshot. Table 2 also suggests that further work should be done to simplify the handling of polarization damping to reduce the computational time further.

Table 2. Timings for the Construction of the ACPE from 400 Acrolein in Water Configurations[a]

	time (s)	% of total time
CG hydrogen	405	1.8
CG oxygen	303	1.3
CG clusters	41	0.16
translations	3	0.011
config. pol.	597	2.3
ACPE pol. cutoffs	5649	21.6
ACPE total	6998	26.7

[a] CG refers to the time to generate the CG sites via K-means++. Config. pol. indicates the time to compute the polarization in the individual configurations, and ACPE pol. cutoffs is the time to identify appropriate polarization cutoffs automatically.

4.2. Solvent Structure in the ACPE Model.

One of the primary motivations underlying the ACPE model is to retain important local structural features of the environment that would not be found in a more traditional implicit solvent model. Continuum solvent models typically have difficulty describing local solute−solvent hydrogen bonding interactions, for instance, which sometimes necessitates the inclusion of explicit solvent molecules. Because ACPE derives its representation of the environment from explicit solvent configurations, it naturally retains some of these localized interactions.

Consider the hydrogen bonding interactions between pyrimidine and solvent water. The contours in Figure 5 plot the distribution of hydrogen atoms near the two hydrogen-bond accepting nitrogen atoms in pyrimidine over the 400 configurations, projected onto the molecular *XY* or *XZ* planes. The red symbols represent hydrogen atom CG sites identified by the K-means++ algorithm. The K-means++ algorithm naturally places CG sites in regions with the highest hydrogen atom density. It captures the variability in hydrogen bond lengths and angles observed across the MD configuration snapshots.

Another perspective on solute−solvent interactions can be gleaned from the radial distribution functions (RDFs), which are plotted in Figure S1 in the Supporting Information. One can compute the average number of hydrogen bonds by counting the number of atoms within the first solvent shell of the N···H−O RDF and dividing by the number of snapshots considered. From the MD simulations, the first solvent shell ends at an N···H distance of 2.7 Å, and integrating the RDF indicates that a pyrimidine nitrogen averages 1.84 hydrogen bonds. For the ACPE model, we obtain the number of hydrogen bonds by counting the number of CG sites within in the same N···H distance and dividing by the scaling factor (5 here). Doing so, one finds an average number of 1.90 hydrogen

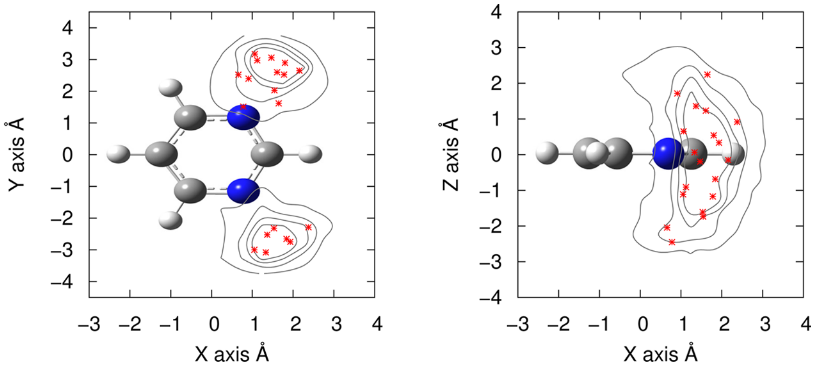


Figure 5. Hydrogen bonding distribution for molecular dynamics configurations and K-means++ generated grid points around pyrimidine.

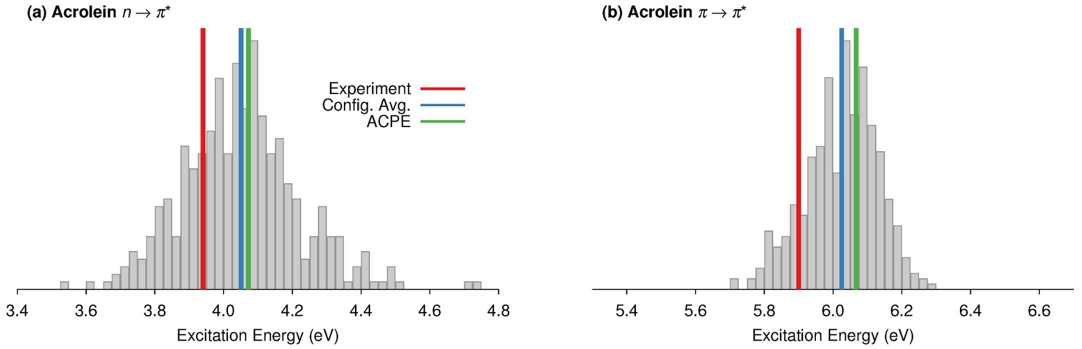


Figure 6. Histograms of the first and second singlet excitation energies of acrolein compared with the experimental, configurational average, and ACPE values. Box heights indicate the number of configurations with this excitation energy.

Table 3. Comparison of CAM-B3LYP/aug-cc-pVTZ Excitation Energies E and Solvatochromic Shifts ΔE for Three Solutes in Aqueous Solution[a]

	acrolein		acetone		pyrimidine	
	$E(n \rightarrow \pi^*)$	ΔE	$E(n \rightarrow \pi^*)$	ΔE	$E(n \rightarrow \pi^*)$	ΔE
gas	3.84		4.49		4.55	
PCM	3.97	0.13	4.59	0.11	4.71	0.16
config. avg.	4.05 ± 0.17	0.21 ± 0.17	4.65 ± 0.16	0.17 ± 0.16	5.01 ± 0.20	0.46 ± 0.20
ACPE	4.07	0.23	4.64	0.16	5.01	0.46
experiment	3.94[b]	0.25[b]	4.68[c]	0.22[c]	4.57,[d] 4.84[e]	0.35,[d] 0.62[f]
	$E(\pi \rightarrow \pi^*)$	ΔE				
gas	6.46					
PCM	5.76	−0.70				
config. avg.	6.03 ± 0.10	−0.44 ± 0.10				
ACPE	6.07	−0.40				
experiment	5.90[b]	−0.52[b]				

[a] The ACPE excitation energies are the average of three calculations. The individual values can be found in the Supporting Information. [b] Ref 67. [c] Ref 71. [d] Ref 72. [e] Ref 73. [f] Inferred using the gas-phase excitation energy from ref 72 and the solution-phase excitation energy from ref 73.

bonds in ACPE, in very good agreement with the explicit MD result.

4.3. Excitation Energies in an Aqueous Environment.

Next, we examine the performance of ACPE for reproducing small-molecule vertical excitation energies in solution. We compare the ACPE excitation energies against values obtained from a traditional configurational average approach, a polarizable continuum model, and experiment. One should bear in mind that discrepancies between the predicted and experimental results can arise for multiple reasons, including: limitations of the TDDFT functional, basis set, and embedding model; the quality of the ensemble generated from OPLS; and the finite cluster truncation of the bulk solvent model in addition to the ACPE approximations. Moreover, experimentally reported excitation energies do not always correspond to the vertical excitation energies computed here. Accordingly, the comparison between the configurational average and the ACPE model results provides more direct insight into the behavior of the ACPE approximations.

Figure 6 plots a histogram of excitation energies from each of the 400 individual polarizable embedding calculations, where the height of each box corresponds to the number of configurations exhibiting excitation energies within the particular energy interval. Across the 400 sampled configurations, the $n \rightarrow \pi^*$ excitation energies occur between 3.55 and 4.74 eV with an average value of 4.05 ± 0.17 eV. This average excitation energy is in good agreement with the value of 4.11 eV from an earlier work using polarizable embedding and the M2P2 force field.[69] The second excitation in acrolein, $\pi \rightarrow \pi^*$, occurs between 5.71 and 6.29 eV across the 400 configurations with an average of 6.03 ± 0.10 eV. The configurational averages for these two excitation energies also lie within 0.1−0.2 eV of experiment (Table 3), which is well within the accuracy expected for TDDFT with CAM-B3LYP.[70] Plots of the excitation energies as a function of the number of solvent configurations averaged over suggest that both of these configurational averages are converged to within a few hundredths of an eV with respect to the number of configurations sampled (see Figure S2 in the Supporting Information).

By definition, the ACPE model does not capture the full distribution of excitation energies observed over the 400 configurations. However, it would ideally mimic the configurational average. Indeed, for both the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ transitions in acrolein, the ACPE reproduces the configurational average to within 0.03−0.05 eV (Figure 6 and Table 3), despite performing only a single QM calculation instead of 400. Though the specific CG sites identified by the K-means clustering algorithm will vary with the initial guess, the resulting excitation energies from three different initial guesses varied by only ±0.01−0.02 eV in both states (see Table S2 in the Supporting Information).

Reliable prediction of solvatochromic shifts is often important when modeling electronic excitations in solution. The electronic characters of the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ transitions in acrolein differ notably, which impacts the solvatochromic

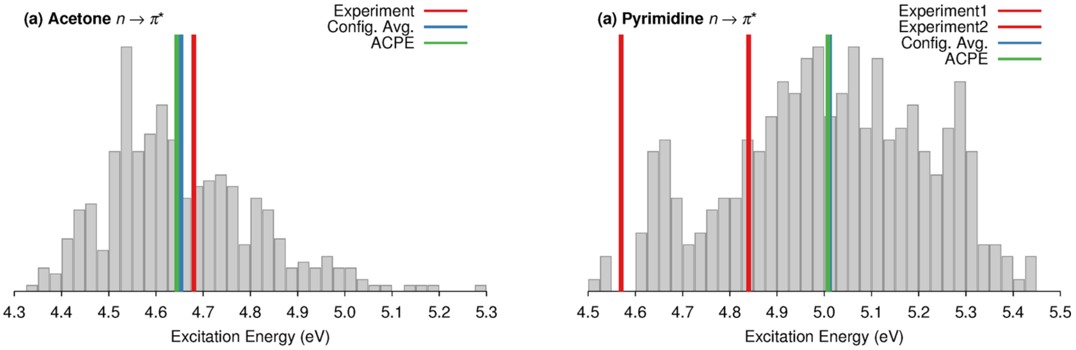


Figure 7. Histograms of the lowest singlet excitation energies of (a) acetone and (b) pyrimidine compared with the experimental, configurational average, and ACPE values. For pyrimidine, two different reported experimental excitation energies are shown.

shifts. The $\pi \rightarrow \pi^*$ excitation shows a sizable red shift of −0.30 eV in iso-octane, while the $n \rightarrow \pi^*$ transition shows much smaller solvatochromic shift in nonpolar solvents.[67] Aidas et al. suggest that the $\pi \rightarrow \pi^*$ shift depends on both electrostatics and intermolecular polarization effects, while the $n \rightarrow \pi^*$ solvent shift is dominated by electrostatic interactions.[67]

Upon switching from the gas phase to an aqueous environment, the ACPE model predicts solvatochromic shifts of 0.23 eV for the $n \rightarrow \pi^*$ transition and −0.40 eV for the $\pi \rightarrow \pi^*$ one (Table 3). These shifts agree within 0.05 eV of the configurational average shifts of 0.21 and −0.44 eV, respectively, which indicates that the ACPE model is reproducing the key features of the configurational average. Agreement between the ACPE and experimental[67] solvatochromic shifts is less uniform but still good with errors of 0.02 and 0.12 eV for the two states, respectively. As demonstrated in our previous work,[30] the coarse-graining of the force field parameters reproduces the original electrostatics well, but it is more approximate for polarization. This helps explain why the performance is worse for the $\pi \rightarrow \pi^*$ state where polarization matters more. There can also be random variations in the ACPE excitation energies of up to a few hundredths of an eV due to the particular coarse graining obtained in a given calculation (see Table 1 and Table S2 in the Supporting Information). Nevertheless, the solvatochromic shifts with ACPE are more accurate than those obtained with a PCM model, as will be described below.

Next, we consider the lowest singlet excitation energies in acetone and pyrimidine. For acetone, the lowest excitation corresponds to the forbidden $n \rightarrow \pi^*$ transition. The energy of this excitation ranges from 4.33 to 5.29 eV over the sampled MD configurations (Figure 7a). Averaging over all 400 configurations produces a configurational average excitation energy of 4.65 ± 0.16 eV, which is in excellent agreement with both the earlier M3P2 force field prediction of 4.75 eV[69] and the experimental value of 4.68 eV. A single ACPE calculation reproduces the configurational average for the first excitation to within 0.01 eV. The solvatochromic shift of 0.16−0.17 eV predicted by both the ACPE model and the configurational average are also in very good agreement with the experimental shift of 0.22 eV (Table 3).[71]

Figure 7b plots the analogous results for the $n \rightarrow \pi^*$ transition in pyrimidine. The excitation energies range 4.52−5.44 eV over the 400 MD configurations, with a configurational average of 5.01 ± 0.20 eV. Once again, the ACPE calculation reproduces the configurational average excitation energy to within 0.01 eV. Experimentally, the $n \rightarrow \pi^*$ excitation is very broad, making it difficult to assign a precise excitation energy. Values ranging from 4.57[72] to 4.84 eV[73] are reported in the literature. Our predictions agree with the latter value to within 0.20 eV. The predicted ACPE and configurational average solvatochromic shifts of 0.46−0.46 eV are also in similarly good agreement with the corresponding experimental value of 0.62 eV (see Table 3). As for acrolein, the variation in the acetone and pyrimidine ACPE excitation energies with the initial random K-means clustering guess is only a few hundredths of an eV (Table S2).

Finally, it is interesting to compare the ACPE results against those obtained with an implicit PCM water model. As shown in Table 3, the excitation energies in the PCM are consistently lower than the configurational average ones for the cases examined here. The PCM excitation energy errors with respect to the experiment are similar to or slightly smaller than those from the configurational averages or the ACPE model. However, given the few tenths of an eV errors typically expected for valence excitation energies with TDDFT,[70] none of the approaches is clearly superior in terms of the excitation energies. On the other hand, the solvatochromic shifts computed with the configurational averages and/or ACPE model are consistently better than those from the PCM model. This is most notable for pyrimidine, for which it has been argued that obtaining reliable solvatochromic shifts requires the inclusion of several explicit waters.[74] The pyrimidine PCM model shift of 0.16 eV (without any explicit solvent molecules) is reasonably close to the 0.35 eV shift from ref 72, but it is much further away from the value of 0.62 eV value inferred from ref 73. The ACPE and configurational average shifts of 0.46 eV lie in between the two experimental values.

Overall, for these simple examples of computing small-molecule excitation energies in aqueous solution, the ACPE model performs very well. A single QM excitation energy calculation embedded in the ACPE reproduces the excitation energies and solvatochromic shifts obtained from a much more expensive configurational average to within a few hundredths of an eV. Of course, PCM and other simple models often can describe these sorts of homogeneous bulk environments well, especially when a small cluster of explicit solvent molecules is included. In the next section, however, we consider a spatially inhomogeneous model that would be much harder to describe with traditional implicit models.

4.4. Solute at the Benzene−Water Interface.

To test the ability of the ACPE model to treat an inhomogeneous environment, we construct a model system consisting of a phenol solute molecule at the interface of liquid benzene and water. This system was chosen because (1) the two solvents exhibit very different polarities and would create an interface with an asymmetric electrostatic environment and (2) the inherent rigidity of the benzene simplifies the treatment of its force field parameters. Figure 8 shows a sample MD configuration of this system and plots the proportion of water and benzene molecules (averaged over all 400 configuration snapshots) as a function of the z-coordinate in the box. The left side of the box is dominated by benzene molecules, while the right side consists mostly of water ones. The phenol molecule resides right at the interface, with the hydroxyl group oriented toward the water region.

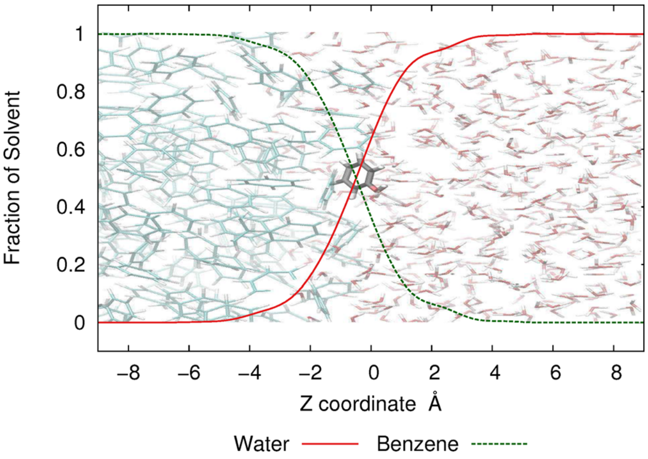


Figure 8. Proportion of water and benzene molecules as a function of the z-coordinate averaged over 400 MD configurational snapshots for phenol at the benzene/water interface.

Figure 9 plots the distribution of excitation energies observed for phenol across the 400 MD configurations. Despite the strong asymmetry of the environment, the ACPE reproduces the configurational average to within less than 0.02 eV (Table 4). Further insight is obtained by investigating the effects of each solvent layer on the phenol first excitation energy separately. Table 4 compares the configurational average and ACPE phenol excitation energies for phenol with only the aqueous solvent molecules present, only the benzene solvent molecules, and in the presence of both solvents. The water-only and benzene-only cases use the same solvent configurations as the system as a whole, just with the other solvent molecules deleted. The two solvents induce opposing solvatochromic shifts on the $S_0 \rightarrow S_1$ excitation. The pure benzene layer red shifts the excitation energy by −0.11 eV, while the pure water layer causes a 0.17 eV blue shift. When both sets of solvent molecules are present, however, the excitation energy undergoes a 0.13 eV blue shift. In other words, the effect of the solvent interface is more than a simple average of the two parts.

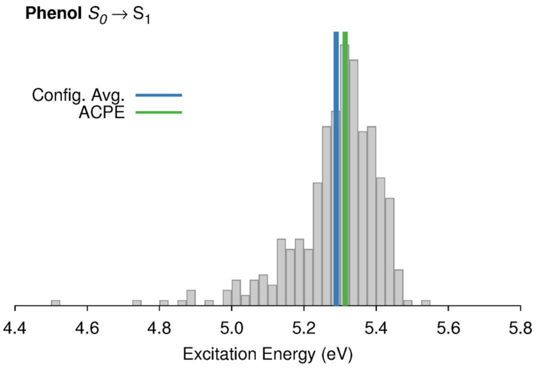


Figure 9. Histogram of excitation energies for phenol lying at the benzene−water interface.

Overall, the water−benzene interface provides a nice example of the robustness of the ACPE model. Despite the heterogeneous and dynamic nature of the environment surrounding the solute, the ACPE model captures the average environment in a single calculation.

5. CONCLUSIONS

In conclusion, we presented an automated procedure for constructing a configuration-averaged condensed-phase environment model around a region of interest based on K-means++ clustering and force field parameter translation procedures. The model has a few adjustable parameters (the number of coarse-graining sites, the initial random seed, and the polarization cutoffs), but fortunately, the results seem relatively insensitive over a range of reasonable choices for these parameter values.

For a species in pure, homogeneous solvent, a PCM model (perhaps augmented with inclusion of a few explicit solvent molecules) offers simplicity, efficiency, and often good-quality predictions. When systems become more complicated and inhomogeneous, however, the ACPE model provides a mechanism to capture the configurational sampling effects in a simplified way. The chief advantages of this approach are that the resulting coarse-grained embedding model (1) retains specific structural features of the underlying atomistic model and (2) it reproduces the conventional configurational average approach with very high accuracy at orders of magnitude lower computational cost. We demonstrated, for example, that it reproduces key locations of hydrogen bonding partners and accurately describes the behavior of a phenol molecule located at the interface of benzene and water solvents. More generally, the ACPE model may prove useful in situations where an inhomogeneous environment precludes the use of more traditional continuum environment models.

Once a set of configurations has been obtained via some sampling procedure, constructing the ACPE model requires

Table 4. Comparison of CAM-B3LYP/aug-cc-pVTZ Excitation Energies E and Solvatochromic Shifts ΔE for Phenol at the Benzene−Water Interface

	water		benzene		interface	
	$E(S_0 \rightarrow S_1)$	ΔE	$E(S_0 \rightarrow S_1)$	ΔE	$E(S_0 \rightarrow S_1)$	ΔE
gas	5.16					
config. avg.	5.33 ± 0.08	0.17	5.05 ± 0.15	−0.11	5.29 ± 0.12	0.13
ACPE	5.33	0.17	5.08	−0.08	5.31	0.16

DOI: 10.1021/acs.jctc.6b00890

minimal computational effort, typically only a fraction of the time required to perform a single embedded excitation energy calculation here. In other words, the computational savings factor for the ACPE model compared to a conventional QM/MM configurational average approaches the number of configurations sampled. At the same time, the ACPE excitation energies reported here all reproduce the configurational average values to within less than 0.1 eV, which is well within the sorts of errors one expects from TDDFT valence excitation energies.

The ACPE model does currently have limitations and opportunities for future work. Most pressingly, all results here utilized a fixed QM region (the frozen solute) in a dynamic environment. In practice, one should also consider the dynamics of the QM region. One possible path forward would be to sample configurations of the QM region, freeze them, and then perform additional sampling of the environment to generate an ACPE for each sampled QM configuration. However, such an approach would neglect explicit couplings between solute and solvent, which can be important.

Additionally, the model is predicated on the notion that the configuration averaging can be performed before the property calculation (e.g., excitation energies) instead of afterward, as is more traditional. This clearly works well in the examples tested here, and it will likely work well in cases where the observable properties of interest occur on time scales which are long relative to the configuration averaging. Experimentally observed nuclear magnetic resonance chemical shifts, for instance, typically represent a time average over nuclear motions. Predictions of observables which occur on much shorter time scales may be less amenable to such a priori configurational averaging.

Finally, the examples here involved rather simple model systems. It will be interesting to extend these ideas to more general systems and a broader range of polarizable force fields. Generalization to more classes of systems might also provide additional insight into how to choose appropriate values for the handful of user-defined parameters in the ACPE model (number of coarse-graining sites, polarization cutoffs, etc.).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.6b00890.

> Detailed mathematical expressions for rotating and translating multipoles and polarizabilities, a radial distribution function for pyrimidine in water, and data on the convergence of the excitation energies (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gregory.beran@ucr.edu; Phone: +1 951 827-7869.

ORCID

Gregory J. O. Beran: 0000-0002-2229-2580

Funding

Funding for this work from the National Science Foundation (Grant CHE-1362465) and supercomputer time from XSEDE (Grant TG-CHE110064) are gratefully acknowledged.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Kelly Theel for helpful discussions.

■ REFERENCES

(1) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1985**, *107*, 154−163.

(2) Stanton, R. V.; Perakyla, M.; Bakowies, D.; Kollman, P. A. *J. Am. Chem. Soc.* **1998**, *120*, 3448−3457.

(3) Zhang, Y.; Liu, H.; Yang, W. *J. Chem. Phys.* **2000**, *112*, 3483−3492.

(4) Ferre, N.; Angyan, J. *Chem. Phys. Lett.* **2002**, *356*, 331−339.

(5) Rod, T. H.; Ryde, U. *J. Chem. Theory Comput.* **2005**, *1*, 1240−1251.

(6) Kastner, J.; Senn, H. M.; Thiel, S.; Otte, N.; Thiel, W. *J. Chem. Theory Comput.* **2006**, *2*, 452−461.

(7) Pulay, P.; Janowski, T. *Int. J. Quantum Chem.* **2009**, *109*, 2113−2120.

(8) Janowski, T.; Wolinski, K.; Pulay, P. *Chem. Phys. Lett.* **2012**, *530*, 1−9.

(9) Nakano, H.; Yamamoto, T. *J. Chem. Theory Comput.* **2013**, *9*, 188−203.

(10) Sodt, A. J.; Mei, Y.; König, G.; Tao, P.; Steele, R. P.; Brooks, B. R.; Shao, Y. *J. Phys. Chem. A* **2015**, *119*, 1511−1523.

(11) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161−2200.

(12) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999−3093.

(13) Mennucci, B. *WIREs Comput. Mol. Sci.* **2012**, *2*, 386−404.

(14) Cram, D. J.; McCarty, J. E. *J. Am. Chem. Soc.* **1954**, *76*, 5740−5745.

(15) Cram, D. J.; Sahyun, M. R. V. *J. Am. Chem. Soc.* **1962**, *84*, 1734−1735.

(16) Sahyun, M. R. V.; Cram, D. J. *J. Am. Chem. Soc.* **1963**, *85*, 1263−1268.

(17) Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2006**, *128*, 6141−6.

(18) Schutz, C. N.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 400−17.

(19) Li, L.; Li, C.; Zhang, Z.; Alexov, E. *J. Chem. Theory Comput.* **2013**, *9*, 2126−2136.

(20) Kukic, P.; Farrell, D.; McIntosh, L. P.; García-Moreno, B.; Jensen, K. S.; Toleikis, Z.; Teilum, K.; Nielsen, J. E. *J. Am. Chem. Soc.* **2013**, *135*, 16968−16976.

(21) An, L.; Wang, Y.; Zhang, N.; Yan, S.; Bax, A.; Yao, L. *J. Am. Chem. Soc.* **2014**, *136*, 12816−9.

(22) Sanchez, M. L.; Aguilar, M. A.; Olivares del Valle, F. J. *J. Comput. Chem.* **1997**, *18*, 313−322.

(23) Sanchez Mendoza, M.; Aguilar, M.; Olivares del Valle, F. *J. Mol. Struct.: THEOCHEM* **1998**, *426*, 181−190.

(24) Sanchez, M. L.; Martin, M. E.; Aguilar, M. A.; Olivares del Valle, F. J. *J. Comput. Chem.* **2000**, *21*, 705−715.

(25) Coutinho, K.; Georg, H.; Fonseca, T.; Ludwig, V.; Canuto, S. *Chem. Phys. Lett.* **2007**, *437*, 148−152.

(26) Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 7821−7826.

(27) Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. *Chem. Rev.* **2015**, *115*, 6312−6356.

(28) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 10095.

(29) Gusarov, S.; Ziegler, T.; Kovalenko, A. *J. Phys. Chem. A* **2006**, *110*, 6083−6090.

(30) Theel, K. L.; Wen, S.; Beran, G. J. O. *J. Chem. Phys.* **2013**, *139*, 081103.

(31) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73*, 325−348.

(32) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227−249.

(33) Warshel, A. *J. Phys. Chem.* **1979**, *83*, 1640−1652.

(34) Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283−422.

(35) Malcolm, N. O. J.; McDouall, J. J. W. *J. Mol. Struct.: THEOCHEM* **1996**, *366*, 1−9.

(36) Florián, J.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 5583−5595.

(37) Galvan, I.; Sanchez, M.; Martin, M.; Olivares del Valle, F.; Aguilar, M. *Comput. Phys. Commun.* **2003**, *155*, 244−259.

(38) Galván, I. F.; Martín, M. E.; Aguilar, M. a. *J. Comput. Chem.* **2004**, *25*, 1227−33.

(39) Berkhin, P. Survey of Clustering Data Mining Techniques. Technical Report, Accrue Software, San Jose (2002). http://www.cc.gatech.edu/isbell/reading/papers/berkhin02survey.pdf.

(40) Inaba, M.; Katoh, N.; Imai, H. *Proceedings of the Tenth Annual Symposium on Computational Geometry* **1994**, 332−339.

(41) Arthur, D.; Vassilvitskii, S. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics Philadelphia* **2007**, 1027−1035.

(42) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Oxford Science Publications: Oxford, 1987.

(43) Sebetci, A.; Beran, G. J. O. *J. Chem. Theory Comput.* **2010**, *6*, 155−167.

(44) Wen, S.; Beran, G. J. O. *J. Chem. Theory Comput.* **2011**, *7*, 3733−3742.

(45) Neill, D. P. O.; Allan, N. L.; Manby, F. R. In *Accurate Quantum Chemistry in the Condensed Phase*; Manby, F., Ed.; CRC Press: Boca Raton, FL, 2010; pp 163−193.

(46) Stone, A. J.; Misquitta, A. J. *Int. Rev. Phys. Chem.* **2007**, *26*, 193−222.

(47) Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233−239.

(48) Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047−1064.

(49) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128−1132.

(50) LeSueur, C. R.; Stone, A. J. *Mol. Phys.* **1993**, *78*, 1267−1291.

(51) Misquitta, A. J.; Stone, A. J. *J. Chem. Theory Comput.* **2008**, *4*, 7−18.

(52) Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 19−32.

(53) Williams, G. J.; Stone, A. J. *J. Chem. Phys.* **2003**, *119*, 4620−4628.

(54) Misquitta, A. J.; Stone, A. J. CamCASP v5.6. http://www-stone.ch.cam.ac.uk/programs.html (accessed February 23, 2011).

(55) Ivanic, J.; Ruedenberg, K. *J. Phys. Chem.* **1996**, *100*, 6342−6347.

(56) Stone, A. J. *The Theory of Intermolecular Forces*; Clarendon Press: Oxford, 2002.

(57) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549−64.

(58) Ren, P.; Wu, C.; Ponder, J. W. *J. Chem. Theory Comput.* **2011**, *7*, 3143−3161.

(59) Olsen, J. M.; Aidas, K.; Kongsted, J. *J. Chem. Theory Comput.* **2010**, *6*, 3721−3734.

(60) Aidas, K.; Angeli, C.; Bak, K. L.; Bakken, V.; Bast, R.; Boman, L.; Christiansen, O.; Cimiraglia, R.; Coriani, S.; Dahle, P.; Dalskov, E. K.; Ekström, U.; Enevoldsen, T.; Eriksen, J. J.; Ettenhuber, P.; Fernández, B.; Ferrighi, L.; Fliegl, H.; Frediani, L.; Hald, K.; Halkier, A.; Hättig, C.; Heiberg, H.; Helgaker, T.; Hennum, A. C.; Hettema, H.; Hjertenaes, E.; Høst, S.; Høyvik, I.-M.; Iozzi, M. F.; Jansík, B.; Jensen, H. J. A.; Jonsson, D.; Jørgensen, P.; Kauczor, J.; Kirpekar, S.; Kjaergaard, T.; Klopper, W.; Knecht, S.; Kobayashi, R.; Koch, H.; Kongsted, J.; Krapp, A.; Kristensen, K.; Ligabue, A.; Lutnaes, O. B.; Melo, J. I.; Mikkelsen, K. V.; Myhre, R. H.; Neiss, C.; Nielsen, C. B.; Norman, P.; Olsen, J.; Olsen, J. M. H.; Osted, A.; Packer, M. J.; Pawlowski, F.; Pedersen, T. B.; Provasi, P. F.; Reine, S.; Rinkevicius, Z.; Ruden, T. A.; Ruud, K.; Rybkin, V. V.; Sałek, P.; Samson, C. C. M.; de Merás, A. S.; Saue, T.; Sauer, S. P. A.; Schimmelpfennig, B.; Sneskov, K.; Steindal, A. H.; Sylvester-Hvid, K. O.; Taylor, P. R.; Teale, A. M.; Tellgren, E. I.; Tew, D. P.; Thorvaldsen, A. J.; Thøgersen, L.; Vahtras, O.; Watson, M. A.; Wilson, D. J. D.; Ziolkowski, M.; Ågren, H. *WIREs: Comput. Mol. Sci.* **2014**, *4*, 269−284.

(61) Ponder, J. W. TINKER v7.1. http://dasher.wustl.edu/tinker/ (accessed August 31, 2011).

(62) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(63) Miertuš, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117−129.

(64) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian Inc.: Wallingford, CT, 2009.

(65) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51−57.

(66) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(67) Aidas, K.; Møgelhøj, A.; Nilsson, E. J. K.; Johnson, M. S.; Mikkelsen, K. V.; Christiansen, O.; Soderhjelm, P.; Kongsted, J. *J. Chem. Phys.* **2008**, *128*, 194503.

(68) Amos, R. D.; Handy, N. C.; Knowles, P. J.; Rice, J. E.; Stone, A. J. *J. Phys. Chem.* **1985**, *89*, 2186−2192.

(69) Schwabe, T.; Olsen, J. M. H.; Sneskov, K.; Kongsted, J.; Christiansen, O. *J. Chem. Theory Comput.* **2011**, *7*, 2209−2217.

(70) Leang, S. S.; Zahariev, F.; Gordon, M. S. *J. Chem. Phys.* **2012**, *136*, 104101.

(71) Renge, I. *J. Phys. Chem. A* **2009**, *113*, 10678−10686.

(72) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Soc.* **1959**, 1240−1246.

(73) Börresen, H. C. *Acta Chem. Scand.* **1963**, *17*, 921−929.

(74) Manzoni, V.; Lyra, M. L.; Gester, R. M.; Coutinho, K.; Canuto, S. *Phys. Chem. Chem. Phys.* **2010**, *12*, 14023−14033.

DOI: 10.1021/acs.jctc.6b00890