JCTC Journal of Chemical Theory and Computation

Toward a Practical Method for Adaptive QM/MM Simulations

Rosa E. Bulo,*,† Bernd Ensing,‡ Jetze Sikkema,† and Lucas Visscher†

Department of Theoretical Chemistry, Vrije Universiteit Amsterdam, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands and Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

Received March 27, 2009

Abstract: We present an accurate adaptive multiscale molecular dynamics method that will enable the detailed study of large molecular systems that mimic experiment. The method treats the reactive regions at the quantum mechanical level and the inactive environment regions at lower levels of accuracy, while at the same time molecules are allowed to flow across the border between active and environment regions. Among many other things, this scheme affords accurate investigation of chemical reactions in solution. A scheme like this ideally fulfills the key criteria applicable to all molecular dynamics simulations: energy conservation and computational efficiency. Approaches that fulfill both criteria can, however, result in complicated potential energy surfaces, creating rapid energy changes when the border between regions is crossed. With the difference-based adaptive solvation potential, a simple approach is introduced that meets the above requirements and reduces fast fluctuations in the potential to a minimum. In cases where none of the current adaptive QM/MM potentials are able to properly describe the system under investigation, we use a continuous force scheme instead, which, while no longer energy conserving, still retains a related conserved quantity along the trajectory. We show that this scheme does not introduce a significant temperature drift on time scales feasible for QM/MM simulations.

I. Introduction

Computational studies play an integral role in many fields of chemistry today. Theoretical investigations are no longer restricted to small organic molecules in the gas phase, but large and complex systems, like solutions, proteins, or solids, are modeled. These complex systems require methods that thoroughly sample configuration space, such as molecular dynamics (MD).[1] For the study of chemical reactions occurring within such large systems, it is advisible to treat changes in the electronic structure explicitly using quantum mechanical models (QM). In this manner the bond breaking and forming processes at the reactive sites can be correctly represented. Since the changes in the electronic structure of a system are usually local, one can describe regions that lie far away from the active site in a simplified manner. A popular approximation is the treatment of those regions in a classical manner using empirical force fields (QM/MM).[2] Treating two regions with different levels of quantum chemical methods (QM/QM), like mixed basis set methods[3] or frozen density embedding (FDE),[4,5] is an alternative option.

A broad range of multiscale methods have been developed, and they have been widely applied to systems with localized active sites. However, problems may occur when one applies these general methods to simulate the time evolution of a multiscale solute−solvent system at the longer time scales that have become accessible in recent years. For example, in a QM/MM simulation of a solution containing both QM and MM solvent molecules, the diffusive nature of the solvent causes QM molecules to move out of the active

* Corresponding author. E-mail: bulo@few.vu.nl.
† Vrije Universiteit Amsterdam.
‡ University of Amsterdam.

10.1021/ct900148e CCC: $40.75 © 2009 American Chemical Society
Published on Web 07/30/2009

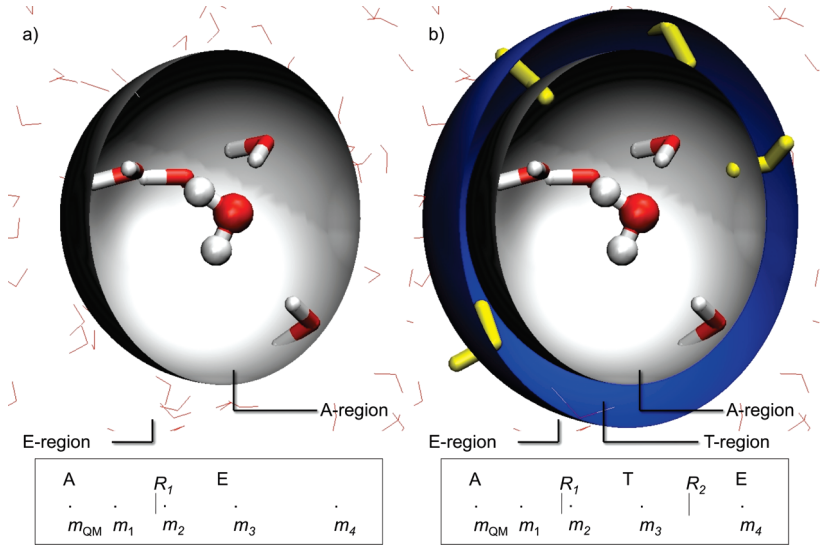


Figure 1. Partitioning of a water box into (a) an A-region (thick water molecules) and E-region (thin lines representing the H_2O) and (b) an A-region, T-region (yellow water molecules) and E-region centered around one water molecule (QM-center: m_{QM}). At the bottom, a schematic 1-dimensional representation of the same partitionings is depicted.

region and MM molecules to move in. This results in a system that treats solvent molecules far away from the active site at a high level of theory, whereas in the sensitive region near the active site, solvent molecules are treated at an inadequate level of theory. To overcome the drawbacks of a rigid partitioning of the system, new methods should be developed that allow the description of diffusive molecules to change 'on the fly' when they cross the border between an active (A) and an environment (E) region.

The simplest version of such a scheme simply chooses an A- and E-region based on a radius around a central active subsystem (Figure 1a). When a solvent molecule crosses this border, its description changes from QM to MM or vice versa. Problematic with this simple scheme is the strong dependence of the absolute potential on the number of atoms that are treated quantum mechanically. Altering the number of QM treated solvent molecules will result in a large and abrupt change in the total energy of the system. Furthermore, because the location of the minima on the QM and MM potential energy surfaces may differ significantly, the system may find itself in a high-energy region after the repartitioning. This effect can lead to sudden large jumps in the forces experienced by the atoms. Indeed, it has been shown that such simulations, even in the NVT ensemble, exhibit extreme accelerations of the atoms.[6] The underlying problem of the simple repartitioning method is the lack of a continuous and smooth potential that defines the forces on the atoms at each time step. Writing the forces as derivatives of such a potential guarantees that Newton's laws apply to the simulations, ensuring reliable statistical data. In practice such simulations are hard to tune because it is difficult to distinguish errors in the setup (such as large time steps) from the inherent flaws of the method.

In order to overcome these problems with the above repartitioning approach, several multiscale methods have been developed which all share a common basis. The system under investigation is no longer sharply divided into an A- and E-region, but contains a transition region (T-region), with inner and outer boundaries R_1 and R_2, (Figure 1b) defined by the distance from the active center. The molecules crossing the T-region gradually change character from QM to MM, allowing the forces to vary in a continuous manner. Among the adaptive QM/MM methods are the 'hot spot' method,[7] ONION-XS,[6] and learn on the fly.[8] These methods combine forces obtained from two QM/MM partitionings, one with and one without the T-region treated at the QM level. They are very efficient, greatly diminish the spurious jumps in the forces, and have resulted in successful applications on a variety of systems.[9] However, for most applications, these methods are not able to fully eliminate discontinuities in the forces. In the absence of a thermostat, this results in strong temperature drifts, and also when a thermostat is present significant deviations from the desired equilibrium situation may occur. Schemes that do entirely eliminate all force discontinuities have been developed for additive pair potential-based (i.e., non-QM) adaptive multiscale setups,[10] in the field of combined atomistic (MM)/coarse grained (CG) simulations.[11]

The problems with the QM/MM methods are rooted in the fact that the molecules in the T-region all acquire a fractional degree of 'QM character' because, unlike combined pair potentials, a QM/MM energy combination can only be defined with integer numbers of molecules in the quantum chemical and classical parts of the system. One effective solution to this problem is to determine at each time step the QM/MM potential energies for all possible partitionings with a different subset of T-region molecules included in the QM calculation. The total potential is then defined as a weighted average of these individual potentials. Recently Heyden et al. introduced such a QM/MM approach, which produced completely continuous forces and was able to conserve the total energy of the system.[12] This approach, however, still has some drawbacks when applied to situations where the QM and the MM energy surfaces exhibit large differences, as will be demonstrated later in this paper.

In this work we present two new developments toward accurate and efficient adaptive QM/MM simulations. In Section II, a new adaptive potential, the difference-based adaptive solvation potential is introduced, which enables energy conserving simulations. Section III compares this potential to the sorted adaptive partitioning potential by Heyden et al. and addresses the advantages of the current approach upon application to chemistry in solution. In Section IV, the common drawback of these two energy conserving methods, the introduction of spurious forces in the T-region, is discussed. A force-based method is introduced as an alternative which, in combination with a bookkeeping algorithm,[10] still retains many of the advantages of a fully Hamiltonian approach. Section V, the Results Section, presents a set of 'proof of principle' simulations and compares results obtained with the two methods discussed in this paper. Finally, Section VI contains our conclusions.

II. Difference-Based Adaptive Solvation

In our adaptive solvation scheme, the targeted fractional 'MM character' of each adaptive molecule (m_i) is captured by a function λ_i that measures the progress of the molecule in the T-region. This function gradually increases from 0 to 1 as a molecule crosses the T-region from the inner border R_1 to the outer border R_2. The progress function applied in this paper, based on the distance r_i of the center of the molecule to the center of the A-region, is presented in eq 1:

$$\lambda_i(r_i) = \begin{cases} 0 & \text{if } r_1 < R \\ \dfrac{(r_i - R_1)^2(3R_2 - R_1 - 2r_i)}{(R_2 - R_1)^3} & \text{if } R_1 \le r_1 \le R_2 \\ 1 & \text{if } r_1 > R_2 \end{cases} \quad (1)$$

The total adaptive potential V^{ad} is defined as a weighted average of QM/MM partitioning energies V_a:

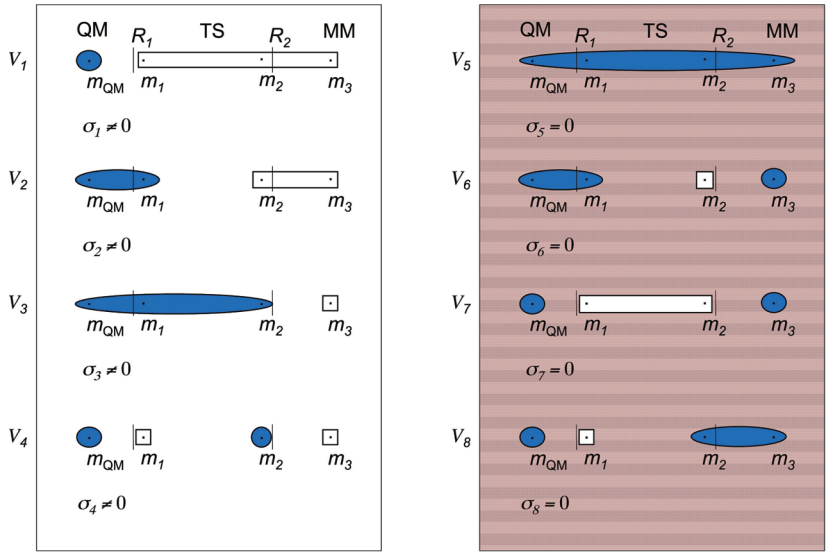
$$V^{ad}(\mathbf{r}) = \sum_a^{2^N} \sigma_a(\mathbf{r}) V_a(\mathbf{r}) \quad (2)$$

in which both the weight σ_a and the potential energy are a function of the coordinates ($\mathbf{r}$) of the N adaptive molecules in the system. The dependence of σ_a on the coordinates will be expressed in terms of the λ_i functions defined above for each of the adaptive molecules. In order to give a detailed description of the method, a few additional definitions are needed. Each QM/MM energy term V_a corresponds to a partitioning of the system into a group of QM molecules (G_a^{QM}) and a group of MM molecules (G_a^{MM}). The sets of λ values for each group will be referred to as $\{\lambda\}_a^{QM}$ and $\{\lambda\}_a^{MM}$.

We can now consider the exact dependence of σ_a on the coordinates. If a total of M molecules is present in the T-region at a given time, then 2^M terms correspond to partitionings that have their QM/MM division line completely inside the T-region. It is those terms that should contribute a non-zero value to the total energy expression of eq 2 (see also Scheme 1). Such a number of energy terms is, however, too large even for the study of small solute molecules. Take for example, the simple test system of one QM water molecule in an MM water liquid. A reasonable approximation would be to define a shell of 4 Å radius as the A-region and surround this by a T-region with a width of 1 Å (Figure 1). At room temperature and pressure, one would then readily find 10 or more water molecules in the T-region, requiring an evaluation of 2^{10} (1 024) relevant QM/MM energy expressions at each time step. At $M = 15$, the number of distinct energy evaluations needed increases to 32 768. Considering that standard MD simulations require several hundred thousand time steps, it is clear that computations such as these will not be feasible.

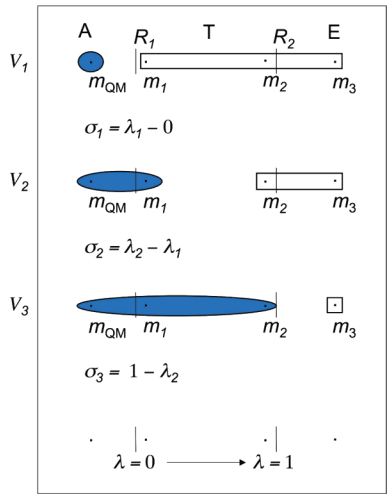
The cause of the unworkable exponential scaling of the intuitive concept described above lies in the inclusion of

Scheme 1. Schematic Representation of the $2^N = 8$ Energy Terms V_a (eq 2) for a System with $N = 3$ Adaptive Molecules[a]



[a] The QM parts of a computation are depicted as blue ellipses, and the MM parts are depicted as white squares. The molecule m_{QM} represents the QM-center and is always computed with a QM method. The terms on the right (red) should have a weight of 0 and will not play a role in the description of the system. Since there are $M = 2$ subsystems in the T-region, $2^M = 4$ terms remain.

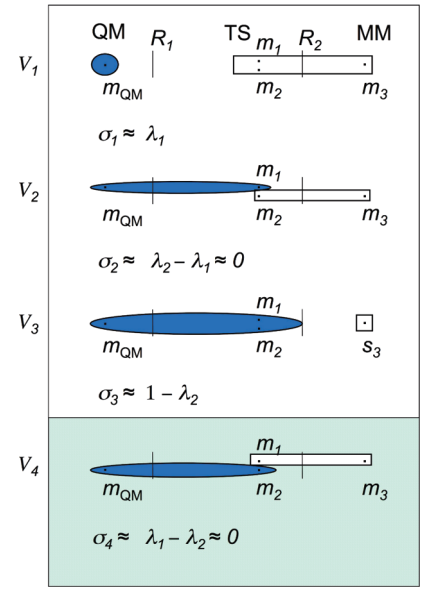
Scheme 2. Schematic Representation of the Contributing Energy Terms V_a for the Difference-Based Adaptive Solvation Method in a System with $N = 3$ Adaptive Molecules



undesirable partitionings that include "almost MM" molecules in the QM description, while "almost QM" molecules are described MM. Ideal σ_a functions should give such energy terms a weight of 0 and focus on the contributions of more "reasonable" partitionings.

In this paper we present a novel potential, which orders the energy terms similarly to the potential by Heyden et al.[12] (Section III) but assigns their weights in a different manner. Both approaches drastically reduce the number of non-zero terms in eq 2 while adhering to a normalization of the weight functions $\sigma_a(\mathbf{r})$. We achieve these goals by first imposing the constraint that energy terms V_a obtain a weight $\sigma_a(\mathbf{r})$ of 0 if there exists a λ in the set $\{\lambda\}_a^{QM}$ that is larger than the smallest λ in the set $\{\lambda\}_a^{MM}$. This amounts to an ordering of the subsystems in the T-region, such that the only contributing partitionings are the following: the term with only the A-region molecules in G^{QM} (V_1 in Scheme 1), the A-region molecules plus the closest T-region molecule in G^{QM} (V_2 in Scheme 1), the A-region molecules plus the two closest T-region molecules in G^{QM} (V_3 in Scheme 1), etc., up to the term in which the entire A- and T-regions are included in G^{QM} (Scheme 2). The exponential scaling of the number of non-zero energy terms is thereby reduced to a linear scaling that is feasible even for systems with large solute molecules. The sorting of energy terms implies that special care needs to be taken in the construction of the weight functions to guarantee the continuity of the potential at instances when two molecules are at comparable distances from the QM center. Suppose we have molecules m_i and m_j with λ_i slightly smaller than λ_j changing their position relative to the QM-center such that λ_i becomes larger than λ_j. At the crossing point of the two λ values, all energy terms that contain m_i in G^{QM} and m_j in G^{MM} cease to contribute, while all previously neglected terms that include m_j in G^{QM} and m_i in G^{MM} become relevant. This could introduce a sudden change that reinvokes the problems encountered in simulations performed without a transition region. We, therefore, need an additional condition: close to degeneracy of the λ values of two molecules, the weights (and the first derivative of the weight

Scheme 3. Schematic Representation of the Contributing Energy Terms V_a for the Difference-Based Adaptive Solvation Method in a System with $N = 3$ Adaptive Molecules[a]



[a] The term V_4 (green) only has a non-zero weight because in the depicted snapshot m_1 and m_2 are at similar distances from the QM-center.

with respect to the molecular coordinates) of the energy terms that assign these molecules to different groups should be zero.

These considerations have led us to define the weight functions σ_a in eq 2 in terms of differences between two individual λ values. These two λ values are the minimum and maximum value for G_a^{MM} and G_a^{QM}, respectively:

$$\sigma_a = \begin{cases} 0 & \text{if } \max(\{\lambda\}_a^{QM}) > \min(\{\lambda\}_a^{MM}) \\ \min(\{\lambda\}_a^{MM}) - \max(\{\lambda\}_a^{QM}) & \text{if } \max(\{\lambda\}_a^{QM}) \leq \min(\{\lambda\}_a^{MM}) \end{cases} \quad (3)$$

Or, completely equivalently, as

$$\sigma_a = \max(\min(\{\lambda\}_a^{MM}) - \max(\{\lambda\}_a^{QM}), 0) \quad (4)$$

For a system with three adaptive molecules, the non-zero σ values take the simple form that is depicted in Scheme 2.

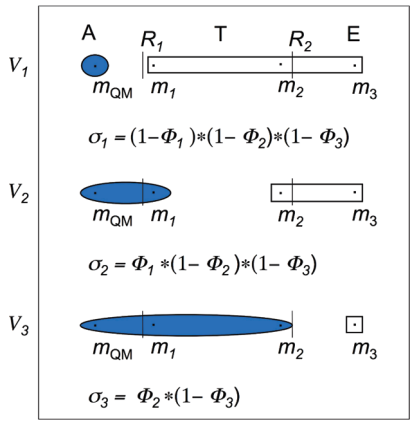
With this simple definition, the weight functions σ_a obey nearly all of the requirements for a smooth and normalized potential.

1) When m_i leaves the T-region at R_1 ($\lambda_i \rightarrow 0$), the sum of weights for the energy terms that include that molecule in G^{QM} equals 1. All weights for the energy terms that contain m_i in G^{MM} become 0. Furthermore, all first derivatives of these weight functions vanish: $\partial\sigma(\mathbf{r})/\partial r_i = 0$.

2) When m_i leaves the T-region at R_2 ($\lambda_i \rightarrow 1$), the sum of weights for the QM/MM terms that contain that molecule in G^{MM} equals 1. All the terms that include the molecule in G^{QM} equal 0. Again, the first derivatives vanish.

3) If m_i exchanges its relative position to the center with m_j, then the energy term V_a with weight $\sigma_a = \lambda_j - \lambda_i$ become

Scheme 4. Schematic Representation of the $M + 1$ Energy Terms V_a with Non-Zero Weights Present in the Total Expression for the Sorted Adaptive Partitioning Potential for a System with M Molecules in the T-Region $(1 - \Phi_3) = 1$



0, and a new energy term V_b with weight $\sigma_b = \lambda_i - \lambda_j$ starts to contribute. Both of these terms equal 0 when $\lambda_i = \lambda_j$, so that the reordering does not cause a discontinuity in the potential. A discontinuity does show up in the derivative of the potential (Figure 2) as a result of the discrete nature of the min()/max() operation.

Because the remaining small discontinuity could affect the dynamics, we investigated a substitution of the discrete min()/max() functions of eq 4 by smooth approximate min()/max() functions. A possible choice for the max() function is the logarithm of a sum of large exponential functions:[13]

$$\max(x_1, ..., x_N) = \frac{1}{\kappa} \ln\left(\sum_{i=1}^{N} e^{\kappa x_i}\right) \tag{5}$$

which can be combined with a similar choice for the min() function (utilizing the fact that $\{\lambda\}_j$ contains only values between 0 and 1):

$$\min(x_1, ..., x_N) = 1 - \frac{1}{\kappa} \ln\left(\sum_{i=1}^{N} e^{\kappa(1-x_i)}\right) \tag{6}$$

The free parameter κ is to be chosen large enough to make small terms in the sum of exponents negligible. These approximate functions smoothen the discontinuity visible in Figure 1 and only give a significant deviation from the true minimum or maximum when two values in the set are similar.

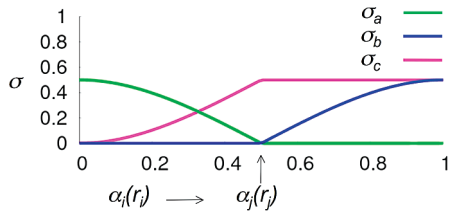


Figure 2. Behavior of the σ functions for a system with only two adaptive subsystems, when m_i moves through a T-region with molecule m_j fixed half-way in the T-region: $\alpha_j(r_j) = 0.5$ with $\alpha_j(r_j) = (r_j - R_1) / (R_2 - R_1)$. σ_a, σ_b, and σ_c correspond to σ_2, σ_4, and σ_1, respectively, in Scheme 3.

A consequence of the introduction of the smooth minimum and maximum functions is a slight increase of the number of partitionings with non-zero weights at instances when two adaptive molecules are at the same distance from the QM center. In this work, we compute energies V_a for which the weight $\sigma_a(\mathbf{r})$ (or its derivatives) exceeds a certain threshold. At most time steps this results in the computation of the minimum of $M + 1$ QM/MM energies (M molecules in the T-region).

While there are to our knowledge currently no faster methods available for doing adaptive QM/MM with fully continuous energies and forces, simulations are still approximately M times slower than a fixed partioning QM/MM simulation. In the target QM/MM simulations, the value M should, thus, be kept small. This is possible because most of the systems of practical interest involve a large active (QM) molecule that contains only a relatively small region where interaction with the surrounding solvent needs to be treated quantum mechanically. Examples are the free coordination space around a transition metal contained in a larger scaffold or the small hydrophilic region in a large hydrophobic (bio)molecule. In these cases, the A-regions may be defined as consisting of one or two atoms on the edge of the active molecule, so that the surrounding T-region will be small enough to keep M manageable. In such cases, the active molecule will have close contacts with MM molecules in the E-region, but this does not affect the scaling. Other ways to keep the costs down will be the development of intelligent restart schemes in which the QM parts of different QM/MM partionings will share their electronic structure information.

III. Sorted Adaptive Partitioning

The difference-based adaptive solvation (DAS) potential is related to the sorted adaptive partitioning potential of Heyden and co-workers.[12] In this sorting approach the same three requirements, as introduced in the previous section, are considered, and for M molecules in the T-region, the total energy is analogously defined as a weighted sum over $M + 1$ terms. The method differs from our setup primarily in the definition of the weights σ_a. Heyden et al. introduce a recursion relation that defines the adaptive potential as a normalized sum of the highest level QM/MM term (V_a, all molecules QM) and another adaptive potential, which excludes the molecule with the least QM character (m_i) from QM description (eq 7). The latter adaptive potential obviously contains terms for all the partitionings of the remaining pool of T-region molecules but can also be written recursively as a weighted average of only two terms, and so on. In contrast to the DAS setup, the expression in eq 7 leads to the definition of the σ from eq 2 as a *product* of the recursive weights Φ_a:

$$V^{\mathrm{ad}}(\mathbf{r}) = V^{\mathrm{ad}(M)}(\mathbf{r}) = \Phi_a(\mathbf{r})V_a(\mathbf{r}) + (1 - \Phi_a(\mathbf{r}))V^{\mathrm{ad}(M-1)}(\mathbf{r}) \tag{7}$$

$$\sigma_a(\mathbf{r}) = \begin{cases} 0 & \text{if } \max(\{\lambda\}_a^{\mathrm{QM}}) > \min(\{\lambda\}_a^{\mathrm{MM}}) \\ \Phi_a(\mathbf{r}) \prod_b (1 - \Phi_b(\mathrm{r})) & \\ \quad b(\max(\{\lambda_b^{\mathrm{QM}}\}) > \max(\{\lambda_a^{\mathrm{QM}}\})) & \text{if } \max(\{\lambda\}_a^{\mathrm{QM}}) \leq \min(\{\lambda\}_a^{\mathrm{MM}}) \end{cases} \quad (8)$$

The auxiliary functions Φ_a can be viewed as progress functions of m_i if this is the solvent molecule in G_a^{QM} with the largest λ value. With $\lambda_i = \max(\{\lambda\}_a^{\mathrm{QM}})$:

$$\Phi_a = \begin{cases} 0 & \text{if } \lambda_i = 1 \\ 1 & \text{if } \lambda_i = 0 \\ (1 + \chi_i)^{-3} & \text{if } 0 < \lambda_i < 1 \end{cases} \quad (9)$$

$$\chi_i = \sum_{j=1}^{i-1} \frac{\lambda_j}{\lambda_i - \lambda_j} + \frac{\lambda_i}{1 - \lambda_i} + \sum_{j=i+1}^{N} \frac{1 - \lambda_j}{\lambda_j - \lambda_i} \lambda_i \quad (10)$$

We note that the smoothing function λ is chosen as a fifth-order spline in the work by Heyden.

Like in DAS, described above, the sorted adaptive partitioning (SAP) potential employs constraints to ensure that "QM character" vanishes if a molecule m_j enters the E-region (eq 9, first line) and the "MM character" vanishes if the molecule enters the A-region. The definition of the functions χ_j (eq 10) ensures the correct behavior at R_1 and R_2. It also ensures vanishing of the contribution of partitionings, assigning molecules at similar distances from the QM-center to different groups.

With this choice of weights, like for the DAS potential, the potential energy V^{ad} is a continuous function of all coordinates. The forces on all atoms are the derivative of this one potential, and the total energy is conserved. As before, the consequence of a solvent molecule m_i passing m_j, while crossing the T-region toward the E-region, is that the weight σ_a of the energy term V_a that treats m_i QM and m_j MM vanishes. Unlike with the DAS potential, it is possible for these weights to have large values just before this situation occurs, leading to a strong variation of the weights in a relatively small region of coordinate space. The effect is especially pronounced in σ_b for the partitioning that treats both molecules MM (Figure 3). The resulting large derivative of the weight functions leads to strong forces on the atoms in the crossing region. A consequence is that commonly chosen time step values in molecular dynamics simulations may no longer be sufficient to conserve the total energy. In fact, as Heyden et al. have shown in an example simulation, the total energy may indeed not be conserved in practice.

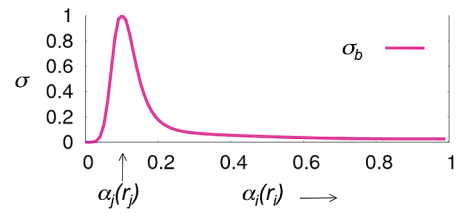


Figure 3. Behavior of a steep σ function for a system with only two adaptive molecules when m_i moves through a T-region with only m_j in it located at $\alpha_j(r_j) = 0.1$ with $\alpha_j(r_j) = (r_j - R_1) / (R_2 - R_1)$. The corresponding partitioning treats both m_i and m_j classically.

IV. Forces in the T-region

In the above, we already mentioned the possibilities for undesirable influences of the chosen weight factors on the computed forces. We did not yet address any problems that can occur when the energy surfaces of the different partitionings differ largely. In order to do so we first consider in more detail the calculation of the forces in simulations with an adaptive potential:

$$V^{\mathrm{ad}}(\mathbf{r}) = \sum_a^{2^N} \sigma_a(\mathbf{r}) V_a(\mathbf{r}) \quad (2)$$

$$F_i = -\frac{\partial V^{\mathrm{ad}}}{\partial r_i} = -\sum_a^{2^N} \sigma_a(\mathbf{r}) \frac{\partial V_a(\mathbf{r})}{\partial r_i} + \frac{\partial \sigma_a(\mathbf{r})}{\partial r_i} V_a(\mathbf{r}) \quad (11)$$

The first term in the force of eq 11 contains simply the weighted force from each contributing partitioning. The second term depends on the derivative of the weight function $\sigma_a(\boldsymbol{r})$ multiplied by the value of the partitioning energy. This introduces a dependence on the relative energies of the different partitionings that is particular to adaptive methods (in simulations with one fixed partitioning force depends only on the slope of the one energy surface). As a consequence, the system may minimize its energy by evolving toward a geometry that maximizes the weight of the partitioning with the lowest absolute energy. This undesirable side effect of adaptive partitioning may be prevented to some extent by aligning the contributing potential energy surfaces as closely as possible. Such a procedure is, however, far from automatic and likely to depend on the particular geometry chosen to define the shifts.

To avoid such errors, it is of interest to consider a deviation from the Hamiltonian approach, in a scheme that interpolates the forces instead of the potential, by using only the first term in eq 11. In contrast to earlier force based methods,[8,9] the weighting functions from the DAS potential guarantee fully continuous forces. This means that many advantages of Hamiltonian approaches can be retained. Integration of the forces over the simulation trajectory should yield a quantity that is conserved throughout the simulation. A so-called bookkeeping term, as used in multiscale atomistic/coarse grain simulations,[10] keeps track of the gain or loss of potential energy throughout the simulation. The conserved quantity is the total energy (using V^{ad} from eq. 2) corrected by this bookkeeping term, and it can be used as a tool to tune other simulation parameters, such as time step and cutoff values.

In the following, we will derive the form of the bookkeeping term mentioned above. In other words, we need to define the corrected potential energy along the path that is associated with the interpolated forces in the first term of eq 11. We can formally add a correction term W^{bk} to eq 2 (eq

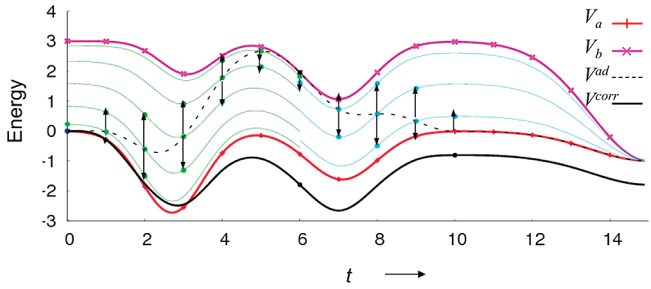
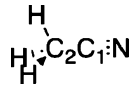


Figure 4. Potential energies along a trajectory. The trajectory follows a loop, with $\mathbf{r}(0) = \mathbf{r}(10)$. V^{ad} is depicted as the thin dashed line. The thick solid line is the potential corrected with the bookkeeping energy contribution.

Table 1. Original and Adjusted Atomic Point Charges in Acetonitrile Force Field



	q original	q modified
N	−0.532	−0.932
C_1	0.481	0.881
C_2	−0.479	−0.479
$(H)_3$	0.177	0.177

12) with a partial derivative that exactly cancels the second term in the forces in eq 11.

$$V^{corr}(t) = V^{ad}(r) + W^{bk}(t) \tag{12}$$

$$F_i^{bk}(\mathbf{r}(t)) = -\frac{\partial W^{bk}(t)}{\partial r_i(t)} = \sum_a^{2^N} \frac{\partial \sigma_a(\mathbf{r})}{\partial r_i} V_a(\mathbf{r}) \tag{13}$$

As an example, we will consider the averaging over two potential energies V_a and V_b, as depicted in Figure 4 as a function of a trajectory with geometry $\mathbf{r}(t)$. We constructed this trajectory such that it forms a loop bringing the system back to its starting coordinates at $t = 10$: ($\mathbf{r}(10) = \mathbf{r}(0)$). The two energy curves have been aligned at an arbitrary point ($t = 15$). The thin dashed line in the figure represents the potential energy V^{ad} as defined in eq 2. Initially the system is in a configuration such that most of the weight is given to V_a, but after a while V_b starts to dominate before the system returns to its initial configuration with V_a again dominant. It is clear from the figure that the shape of the weighted energy curve V^{ad} is mainly determined by the difference in energies of the individual potential energy curves, bearing little resemblance to the shapes of the curves that we aim to average over. With the force-based approach, one utilizes the information gathered during the simulation to define a

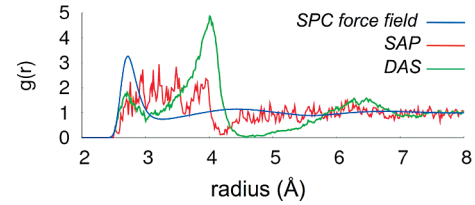


Figure 6. H_2O-OH_2 radial distribution functions around the 'QM'-center for the various methods.

good approximation for the W^{bk} function introduced above. The corrected potential energy curve is depicted in Figure 4 as the black line, and it clearly reflects the shape of the constituent potential energy curves V_a and V_b.

After n time steps along the trajectory, the bookkeeping term W^{bk} can be expressed as the path integral over the force vector from eq 13, as defined in eq 14.

$$\begin{aligned} -W^{bk}(\tau) &= \int_{t=0}^{\tau} F^{bk}(t)\left(\frac{d\mathbf{r}}{dt}\right)dt \\ &= \int_{t=0}^{\tau} \sum_a V_a(t)\left(\vec{\nabla}\sigma_a(t)\left(\frac{d\mathbf{r}}{dt}\right)\right)dt \\ &= \sum_{i=0}^{n} \sum_a V_a(t_i)\left(\frac{\sigma_a(t_{i+1}) - \sigma_a(t_{i-1})}{2\Delta t}\right)\Delta t \\ &= \sum_{i=0}^{n} \sum_a V_a(t_i)\Delta\sigma_a(t_i) \end{aligned} \tag{14}$$

In its discrete form, the bookkeeping term clearly reflects the sum of the accumulated changes in energy at each step along the path. In this manner it corrects for the difference in energy between V_a and V_b, while retaining the different slopes of each of these curves. The value of the correction term is only defined along the path taken by the system during the simulation, and there is no guarantee that the correction term integrates to zero upon revisiting a previous geometry. This feature is also illustrated in Figure 4: at $t = 10$, the geometry is the same as in $t = 0$, but the corrected potential V^{corr} is not. This is due to the fact that the system returns from its largest displacement at $t = 6$ via a route that differs from the initial trajectory. If the system had simply traveled backward via the exact same route, then the corrected energies would have been identical. This reflects how the corrected potential energy V^{corr} depends on the path that is followed and not only on the spatial coordinates of the system.[14] The absence of a single-valued continuous potential in N-dimensional space may result in a small drift of the temperature. However, if this drift is sufficiently small, the simulations will result in a better representation of the system than any other adaptive method.

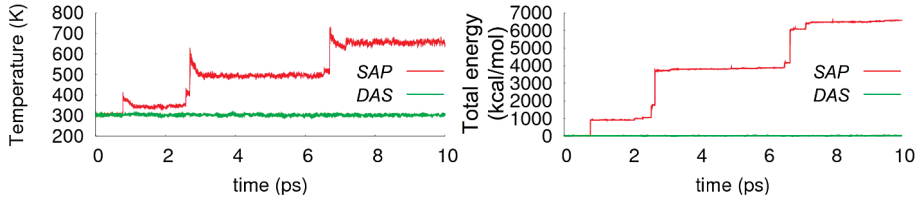


Figure 5. Temperature and total energy during simulation with SAP and DAS potential.

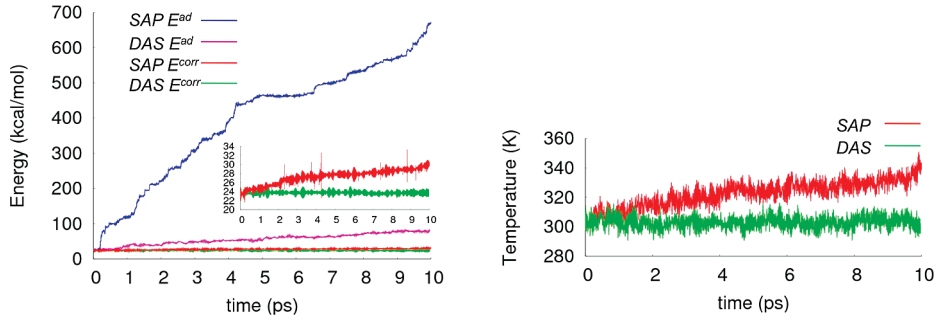


Figure 7. Uncorrected (E^{ad}) and corrected total energies (E^{corr}) obtained with the DAS and SAP interpolated forces and the bookkeeping correction.

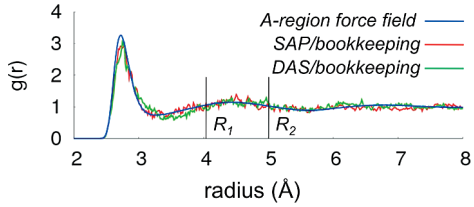


Figure 8. H_2O-OH_2 radial distribution functions obtained with the DAS and SAP interpolated forces.

Table 2. Drift in the Corrected Total Energy for DAS and SAP Force Simulations with Varying Time Steps

	Energy drift in kcal/mol·ps	
ts (fs)	DAS	SAP
0.5	0.017	0.648
0.2	0.003	0.030
0.1	0.022	0.023

V. Results

As an illustrative case study we present NVE simulations for two simple systems, namely water in water and acetonitrile in water. Both systems, containing approximately 3000 water molecules, and around seven water molecules in the active region, can be viewed as representative for target adaptive QM/MM simulations. We used two different classical force fields to represent the two levels of description to provide a quick model for the target QM/MM setup. This also provides us with the possibility to assess the quality of the simulation by comparing the results with the full 'QM' alternative, which would be unfeasible if we had chosen a true QM/MM setup. The NAMD classical molecular dynamics package[15] was used for the energy and force evaluation and the time evolution of the system. For the latter, the program uses a leapfrog algorithm. The composite energies and forces, as described in the previous section, were obtained with a python wrapper script (PyMD) developed for this purpose.

For the MM/MM study of water in water, we started with a box, 30 Å in diameter, containing water that was equilibrated to room temperature and pressure of 400 ps with a time step of 0.5 fs and a flexible TIP3P(Fs) force field.[16] We then evolved the system for 10 ps with a time step of 0.5 fs with both the DAS and SAP potentials in the microcanonical ensemble (NVE). The water in the environment region is described with the same flexible TIP3P(Fs) force field as before, while the water in the active region is

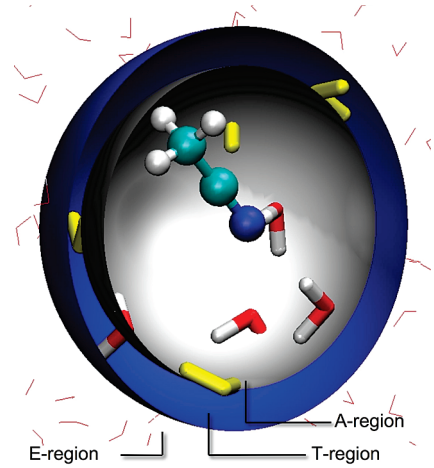


Figure 9. Division of the acetonitrile in water system for adaptive QM/MM.

described with the flexible force field SPC(Fw).[17] The A-region is a sphere with a radius of 4 Å around the oxygen atom of a central water molecule, and the T-region is a shell of 1 Å around the A-region. The 'QM' energy was shifted by the TIP3P(Fs)−SPC(Fw) energy difference for the gas-phase TIP3P(Fs)-optimized solvent molecules, augmented by an additional two-body interaction energy difference, obtained again with the TIP3P(Fs)-optimized coordinates. The aim of the study of acetonitrile in water is to combine an accurate force field in the A-region with a force field in the E-region that is known to fail at a short distance from the solute. As the accurate force field, we chose once again a SPC(Fw) description of water with the AMBER six-center force field for acetonitrile.[18] As the second force field, we chose the TIP3P(Fs) water, in combination with the same AMBER force field for acetonitrile, but this time with a different choice for the atomic charges. We increased the dipole moment of acetonitrile to make the interaction with the solvent artificially bad at short distances. The new partial charges of the acetonitrile atoms are presented in Table 1. The acetonitrile molecule was placed in a water box with a diameter of 30 Å and equilibrated to room temperature and pressure of 400 ps with the flexible TIP3P(Fs) force field. The system was then evolved for 10 ps with a time step of 0.5 fs, using the DAS interpolated forces, with an A-region of 4 Å around the nitrogen atom and a T-region of 1 Å wide.

Water in Water. In order to study the effect of the spurious forces, as discussed in Section IV, we first evolved the system using the full forces from the potential energy

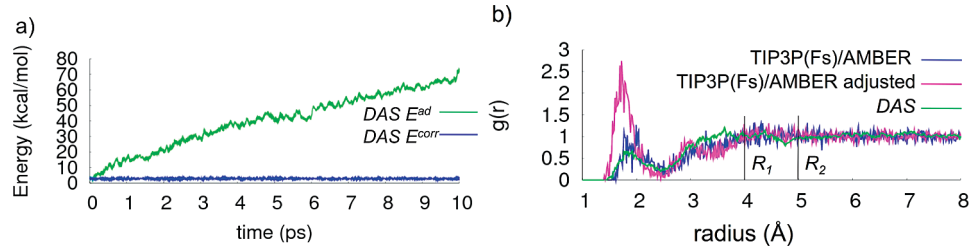


Figure 10. (a) Uncorrected and corrected total energy obtained with the DAS forces. (b) N−H_1 radial distribution functions for a working force field (blue line), a force field that fails at short distances (pink line), and for one obtained with an adaptive mixture of the two in the DAS formulation (green line).

V^{ad}. The adaptive MM/MM simulations with the SAP potential resulted in severe heating of the system in the NVE ensemble. The temperature changes occurred in separate events, corresponding to situations where several water molecules had the same very small λ values. As depicted in Figure 2, this is exactly when very large spurious forces on the atoms can be expected. Simulations with the DAS potential under the exact same circumstances conserved both the temperature and the total energy throughout the run. Figure 5 depicts the behavior of the temperature and total energy throughout the trajectory with both methods.

As expected, the radial distribution functions $g(r)$, obtained from the simulation with the adaptive methods, are not the desired mixture of $g(r)$ from the two force fields used, despite the energy shift that was applied to obtain a reasonable alignment of the potential energy surfaces. We extracted the $g(r)$ from the SAP simulation from the first 1 ps of the simulation where the system still has a temperature of 300 K. Figure 6 shows us how dramatic the error that is introduced for the $g(r)$ between the central water oxygen and all surrounding oxygen atoms. For both methods, the solvent molecules are pushed toward the A-region to maximize the contribution of the partition that describes most molecules with the energetically favorable QM potential. The effect is slightly more pronounced for the DAS result than for that of the SAP result.

We, thus, performed additional simulations using the force-based approach with the weight functions σ_a from the SAP and the DAS methods, respectively, and the bookkeeping correction to the energy, leading to the total energies depicted in Figure 8. The uncorrected total energy obtained with SAP forces displays a strong drift. This is due to the fact that the SAP interpolation is more rapid than that of the DAS approach. The corrected total energy E^{corr} is conserved well only with the DAS forces. Another way of stating this would be to say that, with the SAP forces, the time steps taken in the simulation are too large to give a good approximation of the integral in eq 14. In addition, the DAS approach fully conserves the temperature all throughout the simulation (Figure 7).

Figure 8 shows how the $g(r)$ for both methods are now in perfect agreement with the expected density distribution of water from the combined force fields. With M the number of solvent molecules in the T-region, on average 1.5 times $M + 1$ partitions needs to be computed with the DAS approach, which is an acceptable overhead compared to the number (M + 1) computed in a SAP simulation.

In order to estimate the time step required for a satisfactory performance of the SAP force interpolation, additional short simulations (1 ps) with smaller time steps were performed (Table 2). In all simulations, the bookkeeping correction was applied to the depicted energy. When a time step of 0.2 fs is used, the drift in the bookkeeping quantity with the SAP forces is similarly small to that obtained with a DAS force simulation at a much larger time step.

Acetonitrile in Water. We performed a MM/MM simulation of acetonitrile in water with the DAS forces (Figure 9), applying the bookkeeping correction. Figure 10 depicts the uncorrected and the corrected total energy during the simulation. The figure also shows how the $g(r)$ for (correctly described) acetonitrile in water is reproduced by the adaptive simulation. As expected, the adjusted force field overestimates the number of N−H hydrogen bonds at short distances. This deviation is entirely corrected for in the adaptive formulation.

VI. Conclusions

A new adaptive potential was presented that allows QM/MM molecular dynamics simulations that change the description of solvent molecules on the fly. The scheme defines a potential energy that is smooth enough to allow for a conventionally large time step during the simulation. An interpolated force scheme with fully continuous forces is also introduced, providing correct radial distribution functions and conserving a total quantity along the trajectory. In particular, for large systems of practical interest, these schemes now allow efficient and accurate investigation of processes that involve the dynamical influence of weakly bound solvent molecules.

Acknowledgment. The authors thank The Netherlands Organization for Scientific Research (NWO) for financial support.

References

(1) Frenkel, D.; Smit, B. In *Understanding Molecular Simulation: from Algorithms to Applications*; Academic Press: San Diego, 2002; pp 63−105.

(2) (a) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249. (b) Thole, B. T.; Van Duijnen, P. Th.;. *Theor. Chim. Acta* **1980**, *55*, 307–318. (c) Field, M. J.; Bash, P. A. *J. Comput. Chem.* **1990**, *11*, 700–733. (d) Gao, J. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VHC: New York, 1995; Vol 7, pp 119−185; (e) Sherwood, P. In *Modern*

Methods and Algorithms of Quantum Computing; Grotendorst, J. Ed.; John von Neumann Institute for Computing: Jülich, Germany, 2000; pp 257−277; Carloni, P.; (f) Rothlisberger, U.; Parrinello, M. *Acc. Chem. Res.* **2002**, *35*, 455. (g) Yang, Y.; Yu, H.; York, D.; Elstner, M.; Cui, Q. *J. Chem. Theory Comput.* **2008**, *4*, 2067–2084. (h) Magistrato, A.; DeGrado, W. F.; Laio, A.; Rothlisberger, U.; VandeVondele, J.; Klein, M. L. *J. Phys. Chem. B* **2003**, *107*, 4182–4188.

(3) Moon, S.; Case, D. A. *J. Comput. Chem.* **2006**, *27* (7), 825–836.

(4) (a) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050. (b) Wesolowski, T. A. In *Computational Chemistry: Reviews of Current Trends*; Leszcynski, J., Ed.; World Scientific: Singapore, 2006; Vol*10*, pp 1−82.

(5) (a) Jacob, C. R.; Neugebauer, J.; Jensen, L.; Visscher, L. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2349−2359; (b) Neugebauer, J.; Louwerse, M. J.; Baerends, E. J.; Wesolowski, T. A. *J. Chem. Phys.* **2005**, *122*, 094115; (c) Neugebauer, J.; Jacob, C. R.; Wesolowski, T. A.; Baerends, E. J. *J. Phys. Chem. A* **2005**, *109*, 7805–7814. (d) Neugebauer, J. *J. Chem. Phys.* **2007**, *126*, 134116.

(6) Kerdcharoen, T.; Morokuma, K. *Chem. Phys. Lett.* **2002**, *355*, 257–262.

(7) (a) Kerdcharoen, T.; Liedl, K. R.; Rode, B. M. *Chem. Phys.* **1996**, *211*, 313–323. (b) Hofer, T. S.; Pribil, A. B.; Randolf, B. R.; Rode, B. M. *J. Am. Chem. Soc.* **2005**, *127*, 14231–14238. (c) Schwenk, C. F.; Loeffler, H. H.; Rode, B. M. *J. Am. Chem. Soc.* **2003**, *125*, 1618–1624.

(8) Csányi, G.; Albaret, T.; Payne, M. C.; De Vita, A. *Phys. Rev. Lett.* **2004**, *93* (17), 175503.

(9) (a) Rode, B. M.; Schwenk, C. F.; Tongraar, A. *J. Mol. Liq.* **2004**, *110*, 105–122. (b) Rode, B. M.; Hofer, T. S. *Pure Appl. Chem.* **2006**, *78*, 525–539. (c) Rode, B. M.; Hofer, T. S.; Randolf, B. R.; Schwenk, C. F.; Xenides, D.; Vchirawongkwin, V. *Theor. Chim. Acc.* **2006**, *115*, 77–85.

(10) (a) Ensing, B.; Nielsen, S. O.; Moore, P. B.; Klein, M. L.; Parrinello, M. *J. Chem. Theory Comput.* **2007**, *3*, 1100–1105. (b) Praprotnik, M.; Delle Site, L.; Kremer, K. *J. Chem. Phys.* **2005**, *123*, 224106. (c) Praprotnik, M.; Delle Site, L.; Kremer, K. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2006**, *73*, 066701.

(11) (a) Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. *Phys. Rev. Lett.* **2005**, *95*, 218102. (b) Neri, M.; Anselmi, C.; Carnevale, V.; Vargiu, A. V.; Carloni, P. *J. Phys.: Condens. Matter* **2006**, *18*, S347.

(12) Heyden, A.; Lin, H.; Truhlar, D. G. *J. Phys. Chem. B* **2007**, *111*, 2231–2241.

(13) (a) Donadio, D.; Raiteri, P.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109* (12), 5421–5424. (b) Bulo, R. E.; Donadio, D.; Laio, A.; Molnar, F.; Rieger, J.; Parrinello, M. *Macromolecules* **2007**, *40* (9), 3437–3442.

(14) Delle Site, L. *Phys. Rev. E* **2007**, *76*, 047701.

(15) Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 283–312.

(16) Schmitt, U. W.; Voth, G. A. *J. Chem. Phys.* **1999**, *111*, 9361–9381.

(17) Wu, Y.; Tepper, H. L.; Voth, G. A. *J. Chem. Phys.* **2006**, *124*, 024503.

(18) Grabuledam, X.; Jaime, C.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21* (10), 901–908.

CT900148E