

Electrostatic embedding in large-scale first principles quantum mechanical calculations on biomolecules

Stephen J. Fox, Chris Pittock, Thomas Fox, Christofer S. Tautermann, Noj Malcolm et al.

Citation: *J. Chem. Phys.* **135**, 224107 (2011); doi: 10.1063/1.3665893

View online: <http://dx.doi.org/10.1063/1.3665893>

View Table of Contents: <http://jcp.aip.org/resource/1/JCPSA6/v135/i22>

Published by the [American Institute of Physics](#).

Related Articles

Mode of bindings of zinc oxide nanoparticles to myoglobin and horseradish peroxidase: A spectroscopic investigations

J. Appl. Phys. **110**, 024701 (2011)

A water-swap reaction coordinate for the calculation of absolute protein–ligand binding free energies

JCP: BioChem. Phys. **5**, 02B611 (2011)

A water-swap reaction coordinate for the calculation of absolute protein–ligand binding free energies

J. Chem. Phys. **134**, 054114 (2011)

The axial methionine ligand may control the redox reorganizations in the active site of blue copper proteins

JCP: BioChem. Phys. **4**, 11B601 (2010)

The axial methionine ligand may control the redox reorganizations in the active site of blue copper proteins

J. Chem. Phys. **133**, 175101 (2010)

Additional information on *J. Chem. Phys.*

Journal Homepage: <http://jcp.aip.org/>

Journal Information: http://jcp.aip.org/about/about_the_journal

Top downloads: http://jcp.aip.org/features/most_downloaded

Information for Authors: <http://jcp.aip.org/authors>

ADVERTISEMENT

AIPAdvances

Submit Now

Explore AIP's new
open-access journal

- Article-level metrics now available
- Join the conversation! Rate & comment on articles

Electrostatic embedding in large-scale first principles quantum mechanical calculations on biomolecules

Stephen J. Fox,¹ Chris Pittock,¹ Thomas Fox,² Christofer S. Tautermann,² Noj Malcolm,³ and Chris-Kriton Skylaris^{1,a)}

¹*School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom*

²*Boehringer Ingelheim Pharma GmbH & Co. KG, Lead Identification and Optimization Support, 88397 Biberach, Germany*

³*Accelrys Limited, 334 Cambridge Science Park, Cambridge, CB4 0WN, United Kingdom*

(Received 23 July 2011; accepted 14 November 2011; published online 12 December 2011)

Biomolecular simulations with atomistic detail are often required to describe interactions with chemical accuracy for applications such as the calculation of free energies of binding or chemical reactions in enzymes. Force fields are typically used for this task but these rely on extensive parameterisation which in cases can lead to limited accuracy and transferability, for example for ligands with unusual functional groups. These limitations can be overcome with first principles calculations with methods such as density functional theory (DFT) but at a much higher computational cost. The use of electrostatic embedding can significantly reduce this cost by representing a portion of the simulated system in terms of highly localised charge distributions. These classical charge distributions are electrostatically coupled with the quantum system and represent the effect of the environment in which the quantum system is embedded. In this paper we describe and evaluate such an embedding scheme in which the polarisation of the electronic density by the embedding charges occurs self-consistently during the calculation of the density. We have implemented this scheme in a linear-scaling DFT program as our aim is to treat with DFT entire biomolecules (such as proteins) and large portions of the solvent. We test this approach in the calculation of interaction energies of ligands with biomolecules and solvent and investigate under what conditions these can be obtained with the same level of accuracy as when the entire system is described by DFT, for a variety of neutral and charged species.

© 2011 American Institute of Physics. [doi:10.1063/1.3665893]

I. INTRODUCTION

Properties and processes which involve interactions at the atomic level are ubiquitous in nature. These interactions involve electrons, atoms, and molecules and can in principle only be described by quantum theory. Modern classical force fields which have been extensively parameterised are remarkably accurate and can thus be used instead of quantum calculations in several cases leading to speedups in computational effort which are typically about three orders of magnitude. Nevertheless, force fields can suffer from transferability issues as suitable parameters for new ligands or unusual functional groups are not readily available and may be difficult to determine. A further issue is the inability of force fields to describe electronic charge transfer and polarisation although significant progress in this area is being made via the development of polarisable force fields. Most importantly, force fields cannot be used to describe chemical reactions as the breaking and forming of chemical bonds requires rearrangement of electrons, and the electrons are excluded from force fields. To overcome these limitations, the quantum mechanical/molecular mechanics (QM/MM) approach was introduced by Warshel and Levitt¹ which aims to partition the system in a central active part that is treated by a QM approach and a larger environment which is assumed to be chemically in-

ert and is modelled by a classical force field MM approach. This approach is particularly appealing as the computational effort can be concentrated in the region where high quantum accuracy is needed while the region far from the active site is treated with the much more economic classical force fields. Details of the implementation of such techniques are however highly non-trivial as one needs to carefully define the partitioning of the system in QM and MM parts so that the QM part includes all the chemically important regions and is large enough so that the desired properties are converged with respect to its size. In practice, one has to make an informed compromise between the size and the accuracy with which the quantum region is described so that the cost of the quantum calculations is kept tractable.² An even more difficult choice is how to connect the quantum system to the classical system in the cases where the interface cuts through chemical bonds. Unfortunately this is the rule rather than the exception and a large number of schemes have been developed for connecting the classical and quantum systems. In fact the QM/MM techniques are classified according to the method of creating this interface with variants that are used in chemistry and biochemistry,³⁻⁵ and materials.^{6,7}

In cases where the quantum and classical systems are not connected through chemical bonds, the interaction between the quantum and classical parts includes only non-bonded terms (Coulomb, exchange repulsion, and dispersion). The electrostatic coupling in the QM/MM approach in this

^{a)}Electronic mail: c.skylaris@soton.ac.uk

case is usually referred to as “electrostatic embedding” as the quantum system is embedded in an environment of fixed charges. This is particularly relevant in biomolecular simulations where some molecules (e.g., ligands) may be treated by quantum methods while the surrounding solvent (usually water) is treated by MM.^{8–10} A recent study¹¹ which involved molecular dynamics (MD) simulations of a single “quantum water” embedded into a simulation box filled with “classical waters” showed that the compatibility of the two models depends substantially on the choice of classical force field and extensive testing is required to ensure even qualitative correctness of structural and energetic properties. For example, the TIP5P force field for water when combined with the quantum model produced a qualitatively wrong solvation shell structure with consequent large errors in free energies.

Such inconsistencies are unavoidable and while they can be minimised by the judicious choice of MM and QM approaches, they cannot be eliminated unless a significant part of the solvent is also treated at the quantum level. Performing first principles quantum calculations on entire biomolecular entities including also a significant amount of solvent would result in QM calculations including thousands of atoms which would not be feasible even on modern supercomputers due to the unfavourable (cubic or higher power) scaling of the computational effort with the number of atoms of conventional QM approaches. An alternative approach is that of the frozen-density embedding theory¹² where a high accuracy region is embedded to the frozen density of a lower accuracy region.

Recent developments, based on the quantum mechanical principle of the nearsightedness of electronic matter,¹³ have led to approaches where the computational cost of the calculation increases only linearly with the number of atoms,¹⁴ especially in the area of density functional theory (DFT). Linear-scaling DFT can be used to perform calculations with thousands of atoms hence opening a whole new frontier in accurate simulations. Such calculations have the potential to overcome some of the difficulties encountered by QM/MM approaches by enabling usage of a quantum region which is so large that the method with which the interfacing between QM and MM is done will have negligible effect in the calculated properties as a consequence of the nearsightedness principle. Furthermore, for simulations of many biomolecular systems, linear-scaling DFT calculations can remove altogether the need for coupling the quantum and classical parts through chemical bonds by treating the entire biomolecular assembly at the quantum level and (most of) the solvent classically. In this way, the coupling of the quantum and classical parts can be done purely through electrostatic interactions in a well-defined and unambiguous way.

In Sec. II of this article we introduce the theory of electrostatic embedding in DFT calculations, and we describe its implementation in the ONETEP linear-scaling DFT program¹⁵ and the set up of the simulations we use to test and validate the application of this embedding approach. In Sec. III, we validate the method in the context of calculating interaction energies for a variety of receptor-ligand systems where the role of the receptor is played either by the surrounding solvent (water) or by an entire protein, such as the T4 lysozyme

protein, and the solvent. We finish with our conclusions in Sec. IV.

II. THEORY AND CALCULATION DETAILS

A. Electrostatic embedding in quantum mechanical calculations from first principles

Our aim is to study a quantum system when it is immersed (“embedded”) in an environment of N_{emb} atom-like charge distributions $q_a(\mathbf{r} - \mathbf{R}_a)$, each of which is localised around a point \mathbf{R}_a , and \mathbf{r} is the position in space where the charge distribution is evaluated. The energy of the whole embedded system is therefore composed of the following terms:

$$E_{\text{QM/q}} = E_{\text{QM}} + E_{\text{int}} + E_{\text{q}}, \quad (1)$$

where E_{QM} is the electronic energy of the quantum system (with its density/wavefunctions polarised by the potential due to the embedding charges), E_{int} is the energy of interaction of the electrons and nuclei of the quantum system with the embedding charges, and E_{q} is the electrostatic energy of the embedding charges.

More specifically, in atomic units, we have

$$E_{\text{int}} = \sum_{J=1}^{N_{\text{at}}} \sum_{a=1}^{N_{\text{emb}}} Z_J \int \frac{q_a(\mathbf{r} - \mathbf{R}_a)}{|\mathbf{r} - \mathbf{R}_J|} d\mathbf{r} - \sum_{a=1}^{N_{\text{emb}}} \int \int \frac{q_a(\mathbf{r} - \mathbf{R}_a) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' d\mathbf{r}, \quad (2)$$

where the first term on the right hand side is the Coulomb (i.e., electrostatic) energy of interaction between N_{at} nuclei (with atomic number Z_J) and the N_{emb} embedding charges q_a and the second term is the Coulomb energy of interaction between the electronic density $n(\mathbf{r})$ and the embedding charges. We also have

$$E_{\text{q}} = \sum_{a=1}^{N_{\text{emb}}} \sum_{b>a}^{N_{\text{emb}}} \int \int \frac{q_a(\mathbf{r} - \mathbf{R}_a) q_b(\mathbf{r}' - \mathbf{R}_b)}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' d\mathbf{r}, \quad (3)$$

which is the energy of interaction between the point charges.

The above description is valid whether the embedding has been done either as an *a posteriori* correction after the quantum calculation (and the electronic density is kept “frozen” to its form for the non-embedded quantum system) or when the embedding charges have been present throughout the quantum calculation (by a self-consistent field (SCF) approach) and as a result the electronic density has been polarised accordingly. Here we are interested in the latter case, as applied in DFT calculations. For this self-consistent embedding to take place, the usual Kohn-Sham electronic Hamiltonian,

$$\hat{H}_{\text{KS}} = \hat{T} + \hat{V}_{\text{H}} + \hat{V}_{\text{xc}} + \hat{V}_{\text{ext}}, \quad (4)$$

where the \hat{T} is the electronic kinetic energy operator, \hat{V}_{H} is the Hartree (Coulomb) potential, \hat{V}_{xc} is the exchange-correlation potential, and \hat{V}_{ext} is the external potential, needs to be augmented by a further term due to the potential that each electron will experience from the embedding charge distributions. Therefore, in the presence of the embedding charges

the Hamiltonian becomes

$$\hat{H}_{KS/q} = \hat{T} + \hat{V}_H + \hat{V}_{xc} + \hat{V}_{ext} + \hat{V}_{emb}, \quad (5)$$

where

$$\hat{V}_{emb}(\mathbf{r}) = \sum_{a=1}^{N_{emb}} \int \frac{q_a(\mathbf{r}' - \mathbf{R}_a)}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' = \sum_{a=1}^{N_{emb}} \hat{v}_{emb}^{(a)}(\mathbf{r} - \mathbf{R}_a), \quad (6)$$

with $\hat{v}_{emb}^{(a)}(\mathbf{r} - \mathbf{R}_a)$ being the potential due to each charge distribution $q_a(\mathbf{r}' - \mathbf{R}_a)$, as defined by the above equation.

B. Electrostatic embedding in the ONETEP program

The ONETEP program¹⁵ is based on the reformulation of DFT in terms of the one-particle density matrix,

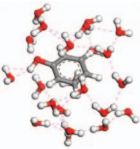
$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^N f_i \psi_i(\mathbf{r}) \psi_i(\mathbf{r}'), \quad (7)$$

where N is the total number of Kohn-Sham molecular orbitals $\{\psi_i(\mathbf{r})\}_{i=1}^N$ and f_i are their occupancies. The one-particle density matrix is the basis of many linear-scaling DFT approaches¹⁴ where the memory and central processing unit (CPU) requirements increase linearly with N_{at} . This is achieved by taking advantage of the exponential decay of the density matrix in systems with a bandgap, which is a manifestation of the ‘‘nearsightedness of electronic matter.’’¹³ In ONETEP the density matrix is expressed in the following form:

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha} \sum_{\beta} \phi_{\alpha}(\mathbf{r}) \mathbf{K}^{\alpha\beta} \phi_{\beta}(\mathbf{r}'), \quad (8)$$

where the ‘‘density kernel’’ \mathbf{K} is the density matrix expressed in the duals of the set of non-orthogonal generalised Wannier functions (NGWFs) (Ref. 16) $\{\phi_{\alpha}(\mathbf{r})\}$. The NGWFs are constrained to be strictly localised within spherical regions centred on atoms and their shape is optimised self-consistently by expressing them in a psinc basis set.¹⁷ This is mathematically equivalent to a plane wave basis set and is therefore systematically improvable to near-complete basis set accuracy¹⁸ without suffering from basis set superposition error (BSSE),¹⁹ characteristic of basis sets which move with the atoms. We demonstrate this point with a set of benchmark calculations in Table I where we present binding energies of a phenol molecule to a configuration of water molecules from its first solvation sphere with ONETEP and with a Gaussian basis set approach (as implemented in the NWChem program²⁰). A kinetic energy cutoff of 800 eV was used to define the psinc basis set. As we can see, with this basis set the binding energies converge rapidly with increasing NGWF radius to the values that are obtained with very large Gaussian basis sets such as the cc-pVTZ. The cc-pVTZ basis set results in 1765 contracted Gaussian basis functions for this system, and consequently matrices of these dimensions, as compared to the 166 NGWFs that are needed in the ONETEP calculations. We should note that the ONETEP NGWFs were initialised to STO-3G contracted Gaussians (excluding the 1s functions for the second row elements, as we are using pseudopotentials), so the effect of their *in situ* optimisation is evident from the results in the table.

TABLE I. Binding energy (BE) of a phenol molecule to its first solvation shell (consisting of 22 water molecules), for the structure shown on the left. Top panel: Energies obtained with ONETEP calculations, with increasing NGWF radii. Bottom panel: Energies obtained with Gaussian basis sets of increasing size with the NWChem program, without and with counterpoise correction²¹ for basis set superposition error. N_{NGWFs} is the number of NGWFs and N_{CGs} is the number of contracted Gaussian functions.

	NGWF radii (Å)	N_{NGWFs}	BE (kcal/mol)
	2.9	166	-11.93
	3.2	166	-12.86
	3.7	166	-8.25
	4.2	166	-7.06
	4.8	166	-7.04
Basis set	N_{CGs}	BE (kcal/mol)	BE with CP (kcal/mol)
STO-3G	195	-23.17	-7.98
3-21G	361	-46.48	-12.55
6-31G*	535	-27.77	-8.95
6-311+G*	817	-17.71	-8.79
6-311++G**	1017	-12.49	-7.39
cc-pVDZ	685	-33.26	-7.28
cc-pVTZ	1765	-19.59	-7.04
cc-pVQZ	3780	-12.41	-7.22

The localisation of the density matrix is effected via the localisation of the NGWFs and truncation of the density kernel elements which correspond to NGWF centres longer than a threshold value r_K . ONETEP has been implemented with advanced parallel algorithms,^{22,23} following the distributed memory paradigm using the message passing interface (MPI) library, and efficient and robust linear-scaling energy minimisation approaches.²⁴ Even though it is a new code, it is already being used in an increasing number of studies of nanostructures²⁵ and biomolecular systems.²⁶⁻²⁸ An example of the linear-scaling behaviour of the code is provided in Figure 1 where the time to perform single-point energy calculations on protein fragments of increasing size is plotted as a function of the number of atoms. In these calculations a value of $r_K = 10.6$ Å was used.

ONETEP obeys periodic boundary conditions, which are naturally compatible with the plane waves from which the psinc basis functions are constructed. As in conventional pseudopotential plane wave approaches,³⁰ certain quantities are generated in reciprocal space. However, these need to be Fourier transforms to real space as this is where the localisation of the density matrix is performed, and for these Fourier transforms the fast Fourier transform (FFT) box approach³¹ is employed, in order to retain the linear-scaling behaviour. The external potential due to the ionic cores which are represented by norm-conserving pseudopotentials is expressed in the Kleinman and Bylander representation,

$$\begin{aligned} \hat{V}_{ext}(\mathbf{r}) &= \hat{V}_{ext,loc}(\mathbf{r}) + \hat{V}_{ext,nl}(\mathbf{r}) \\ &= \sum_{p=1}^{N_{at}} [\hat{v}_{ps,loc}^p(\mathbf{r} - \mathbf{R}_p) + \hat{v}_{ps,nl}(\mathbf{r} - \mathbf{R}_p)], \quad (9) \end{aligned}$$

where N_{at} is the total number of atoms and \mathbf{R}_p is the position of atom p . The potential due to the embedding charges is of

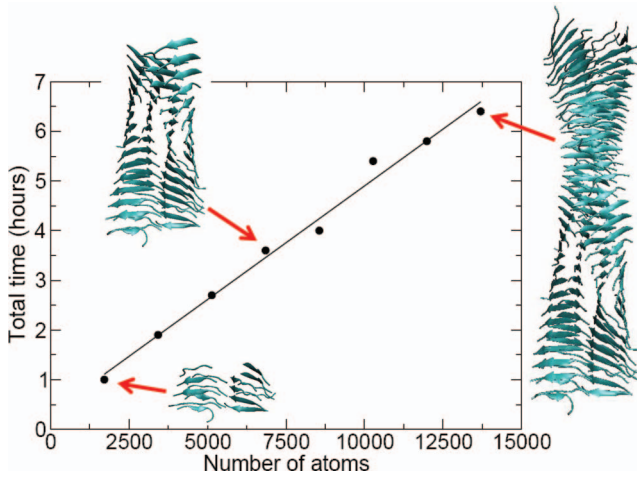


FIG. 1. Scaling of CPU time for single-point energy calculations with ONETEP on segments of amyloid fibrils of increasing size. The calculations were performed on 256 cores of the Iridis3 supercomputer of the University of Southampton and the structures were obtained by kind permission from the authors of Ref. 29.

similar form to the local part of the external potential,

$$\hat{V}_{\text{emb}} = \sum_{a=1}^{N_{\text{emb}}} \hat{v}_{\text{emb}}^{(a)}(\mathbf{r} - \mathbf{R}_a). \quad (10)$$

We therefore generate directly the sum of the two, which can be written in the following form:

$$\begin{aligned} \hat{V}_{\text{ext,loc}}(\mathbf{r}) + \hat{V}_{\text{emb}}(\mathbf{r}) = & \sum_{k=1}^{N_{\text{species}}} \sum_{l=1}^{N_k} \hat{v}_{\text{ps,loc}}^{(k)}(\mathbf{r} - \mathbf{R}_{k,l}) \\ & + \sum_{j=1}^{N_{\text{emb-species}}} \sum_{L=1}^{N_j} \hat{v}_{\text{emb}}^{(j)}(\mathbf{r} - \mathbf{R}_{j,L}), \quad (11) \end{aligned}$$

where $\hat{v}_{\text{ps,loc}}^{(k)}(\mathbf{r} - \mathbf{R}_{k,l})$ is the local pseudopotential for a particular “species” of atomic core (e.g., oxygen) which is centred at position $\mathbf{R}_{k,l}$. In the same manner, $\hat{v}_{\text{emb}}^{(j)}(\mathbf{r} - \mathbf{R}_{j,L})$ is the electrostatic potential due to a particular type of embedding charge distribution which is centred at position $\mathbf{R}_{j,L}$. If the Fourier transform of the potential of each species is provided, the Fourier transform of the total local potential can be obtained as follows:

$$\begin{aligned} \tilde{V}_{\text{ext,loc}}(\mathbf{g}) + \tilde{V}_{\text{emb}}(\mathbf{g}) = & \sum_{j=1}^{N_{\text{species}}} \tilde{v}_{\text{ps,loc}}^{(j)}(\mathbf{g}) \sum_{J=1}^{N_j} e^{-i\mathbf{g}\cdot\mathbf{R}_{j,J}} \\ & + \sum_{p=1}^{N_{\text{emb-species}}} \tilde{v}_{\text{emb}}^{(p)}(\mathbf{g}) \sum_{P=1}^{N_p} e^{-i\mathbf{g}\cdot\mathbf{R}_{p,P}} \\ = & \sum_{j=1}^{N_{\text{species}}} \tilde{v}_{\text{ps,loc}}^{(j)}(\mathbf{g}) S_{\text{ps}}^{(j)}(\mathbf{g}) \\ & + \sum_{p=1}^{N_{\text{emb-species}}} \tilde{v}_{\text{emb}}^{(p)}(\mathbf{g}) S_{\text{emb}}^{(p)}(\mathbf{g}), \quad (12) \end{aligned}$$

where the terms $S_{\text{ps}}^{(j)}(\mathbf{g})$ and $S_{\text{emb}}^{(p)}(\mathbf{g})$ as defined by the above equation are the structure factors³² for each species of pseu-

dopotential and embedding potential, respectively. Therefore the incorporation of the embedding potentials to the Kohn-Sham Hamiltonian can be done with minimal additional cost by building them into the Fourier transform of the local part of the external potential. Furthermore, the above form gives us the flexibility to use any functional form for the embedding charge distribution $q_{\text{emb}}^{(p)}(\mathbf{r})$, since if its Fourier transform $\tilde{q}_{\text{emb}}^{(p)}(\mathbf{g})$ can be obtained, it is possible to obtain an expression for its potential, $\tilde{v}_{\text{emb}}^{(p)}(\mathbf{g})$. The incorporation of the embedding potentials into the electronic Hamiltonian through Eq. (12) ensures that the second term in Eq. (2) is obtained as part of the interaction of the electrons with the external potential (now augmented by the embedding potentials).

The Fourier transform of the embedding potential $\tilde{v}_{\text{emb}}^{(p)}(\mathbf{g})$ is constructed from the Fourier transform for the charge distribution $\tilde{q}_p(\mathbf{g})$ as a solution of the Poisson equation,

$$\tilde{v}_{\text{emb}}^{(p)}(\mathbf{g}) = \frac{4\pi}{\Omega} \frac{\tilde{q}_p(\mathbf{g})}{g^2}. \quad (13)$$

This equation is well-defined, except for $g = 0$. This is a consequence of the fact that the electrostatic potential and energy are divergent for periodically repeated charge distributions with non-zero total charge. We overcome this obstacle by subtracting a uniform charge distribution of total charge equal and opposite to $q_p(\mathbf{r})$ to obtain $\tilde{v}_{\text{emb}}^{(p)}(0)$ as follows:

$$\tilde{v}_{\text{emb}}^{(p)}(0) = \lim_{g \rightarrow 0} \frac{4\pi}{\Omega} \left[\frac{\tilde{q}_p(\mathbf{g})}{g^2} - \frac{\tilde{q}_p(\mathbf{0})}{g^2} \right]. \quad (14)$$

To further simplify this expression we need to provide the explicit form for the charge distribution $\tilde{q}_p(\mathbf{g})$. In this work we consider two kinds of charge distribution,

1. Point charge. In this case $\tilde{q}_p(\mathbf{g}) = \tilde{q}_p(\mathbf{0})$ and $\tilde{v}_{\text{emb}}^{(p)}(0) = 0$.
2. A Gaussian, so that $q_p(\mathbf{r}) = Z_p(\alpha_p/\pi)^{3/2} e^{-\alpha_p r^2}$. In this case $\tilde{q}_p(\mathbf{g}) = Z_p e^{-g^2/(4\alpha_p)}$ and $\tilde{v}_{\text{emb}}^{(p)}(0) = -Z_p \pi / (\Omega \alpha_p)$.

The first term of Eq. (2) and the E_q energy of Eq. (3) are obtained by treating the embedding charge distributions as point charges and including these in the calculation of the nucleus-nucleus interaction energy through an Ewald summation.³³ This is exact under the assumption of no overlap between ionic cores and charge distributions, which applies in practice.

All the terms containing electrostatic energies in Eq. (1) are computed for periodically repeated charge distributions and uniform background charges have to be used as in Eq. (14) to ensure charge neutrality in each step. In cases where the total charge in the unit cell is non-zero, the uniform background charge neutralises the total charge to zero.

Finally, the DFT+D approach as implemented in ONETEP (Ref. 34) is used to include dispersion interactions. This is a purely empirical correction that is added to the DFT energy at the end of the calculation, and as such it does not affect the self-consistent electronic energy optimisation process in any way. Such approaches have been shown to capture dispersion interactions extremely well.³⁵ For example, for the approach we use here with “damping function 1,” with parameters fitted specifically for the PBE exchange-correlation

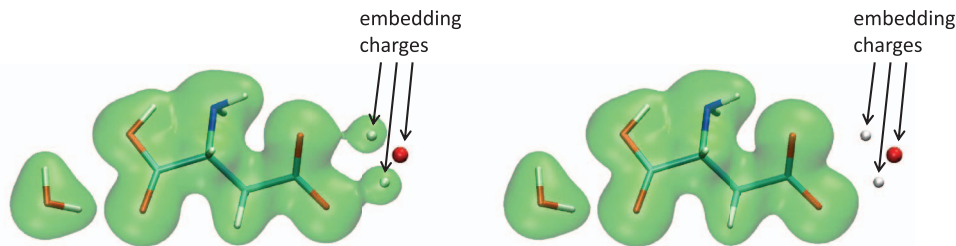


FIG. 2. Electron density isosurfaces of the diaqua ASP complex, with the molecular structure overlaid, for the case $q_0 = -4.0 e$ (isosurface value = $0.2 e a_0^{-3}$). Spilling of electronic density occurs when the embedding is done with point charges (left), but not when Gaussian charge distributions are used (right).

functional, the root mean square (rms) error in the binding energies of a validation set of 21 complexes with a variety of interactions is reduced from 3.6 kcal/mol for pure PBE to about 1 kcal/mol for PBE+D.³⁴

C. Investigation of “charge spilling”

Previous workers have reported that “charge spilling” in electrostatic embedding approaches, which is caused by the attraction of the electronic density to the positive embedding charge distributions⁶ can be a serious issue. This can be exacerbated in cases of large Gaussian basis sets or extended uniform basis sets such as plane waves. To avoid this unphysical effect, the $\hat{v}_{\text{emb}}(\mathbf{r})$ is not the potential of a point charge but it has been given a more diffuse form similar to that of an empirical pseudopotential,³⁶ or the potential of a Gaussian function.³⁷ In order to investigate the possibility of charge spilling in our implementation, we have carried out tests on an aspartate anion (ASP). This is a particularly challenging system as it is negatively charged and is thus expected to have a strong tendency to donate some of its electronic density to neighbouring positive embedding charges. The system studied is shown in Figure 2, where the aspartate anion is hydrogen bonded to two water molecules, one at each of its carboxylic groups. In our tests we represented the atoms of the water molecule bound to the side chain carboxylate group as embedding charges, whilst the rest of the system was treated using DFT.

The energy of interaction was calculated as the difference in the energy of this diaqua complex from its energy when the “quantum” water has been translated 4 Å away from the complex and so that it is effectively no longer hydrogen bonded. These tests were carried out with the embedding charges being represented as point charges and as Gaussians, with a halfwidth of $1.0 a_0$. NGWF radii of 3.7 Å were used. As a result, most NGWF regions overlap with the classical charges so the likelihood of density spilling to occur would be expected to be comparable to that of a plane wave basis set approach. Embedding charges of $q_O = -0.834 e$ for oxygen and $q_H = 0.417 e$ for hydrogen as in the TIP3P force field were used as well as some excessively large values such as $q_O = -2.0 e$ and $-4.0 e$. Careful investigation of the electronic density for the embedding charges of $q_O = -0.834 e$ and $q_H = 0.417 e$, revealed no signs of charge spilling, regardless of the representation as Gaussians or as point charges. Only with the values $q_O = -2.0 e$ and $-4.0 e$, we observed some charge spilling for the case of the point charges but not for the case

of the Gaussian functions. This is demonstrated by examining the electron density isosurfaces, which are shown in Figure 2.

The interaction energies obtained for each set of charges and type (point charge or Gaussian) are presented in Table II. Using the fully quantum calculation as the benchmark, both the point charge and Gaussian methods still agree well with each other and the benchmark calculation with the charge value of $q_0 = -2.0 e$. However, at the even larger embedding charge of $q_0 = -4.0 e$, the point charge model shows large deviation in the energies obtained. As the charges we use for our embedding calculations here are those of the TIP3P water model, it is very unlikely that charge spilling can occur with either representation of the embedding charges. Nevertheless, in our implementation we have selected to use the safer option of the Gaussian smeared embedding charges (but with a halfwidth of $0.3 a_0$ in the calculations that follow, in order to ensure the validity of the assumption of non-overlapping ionic cores of Sec. II B) as the computational effort using the formalism of Sec. II B is the same regardless of the shape of the embedding charge. Also, as the embedding charges are incorporated in the calculation as a component of the external potential (Eq. (12)), they only contribute to a slight increase in the one-off cost of the initialisation of the external potential, which itself is insignificant as it typically takes about 0.001% of the total calculation time. As a result, the inclusion of electrostatic embedding charges results in no observable increase in the cost of the calculation.

D. Molecular dynamics simulations

The electrostatic embedding method we describe in this paper is intended for a variety of applications, such as with rigorous statistical mechanics approaches for the calculation

TABLE II. Interaction energy of the aspartate anion with the quantum water using different partial charges and charge distribution methods for the embedding charges. Charges are given in atomic units (e) and energies in kcal/mol.

Method	q_0	ΔE
All QM	...	-8.4
Point charges	-0.834	-8.5
	-2.0	-7.9
	-4.0	-5.0
Gaussians	-0.834	-8.5
	-2.0	-8.0
	-4.0	-7.2

of free energies of binding as outlined in Sec. III A. Therefore, rather than testing it on simplified model systems, we have sought to test it on physically meaningful structures, free of steric clashes, as the ones that would be obtained in a real-world application scenario via well-equilibrated molecular dynamics simulations.

For the protein ligand complexes, the x-ray crystal structures were checked and protonated with the MOE program,³⁸ then solvated with explicit water in a rectangular box with periodic boundary conditions in the AMBER version 10³⁹ package. Prior to a production MD simulation, an equilibration stage is required since the initial structures and velocities are typically far from the equilibrium phase space of the simulation conditions. To achieve this within the limited time scales that dynamics can be run, which are of the order of ns, complex multiple step equilibration protocols need to be used. The following equilibration procedure was employed: the hydrogens were relaxed keeping all heavy atoms fixed with harmonic restraints in the protein and solvent, then the solvent was relaxed with the protein atoms still fixed. The system was heated gradually to 300 K while still restraining the protein for 200 ps with the NVT ensemble and ran for a further 200 ps with the NPT ensemble at 300 K. This was cooled to 100 K over 100 ps and a series of minimisations was carried out reducing the restraints on the protein heavy atoms in stages (500, 100, 50, 20, 10, 5, 2, 1, and 0.5 kcal mol⁻¹ Å⁻²). Finally, the system was heated to 300 K with no restraints over 200 ps and then ran for a further 200 ps at 300 K with NPT, at the end of which the energy and the density of water in the simulation cell were stabilised and so was the internal structure of the protein as measured by the root mean squared deviation of the backbone atoms from the starting structure, which was 0.75 Å. Production simulations were run for 10 ns with the NVT ensemble at 300 K.

To equilibrate the ligand in a waterbox, the system was heated to 300 K with the NVT ensemble over 300 ps then switched to the NPT ensemble for 200 ps, in order to adjust the volume of the simulation cell and consequently the density of the water. Then the equilibration was completed with the NVT ensemble for 200 ps again at 300 K. The production calculation was with NVT at 300 K for 1 ns.

All MD simulations used the Langevin thermostat, the particle mesh Ewald sum for the long range electrostatics, and a time-step of 2 fs with the SHAKE algorithm. The AM1-BCC method was used to obtain partial charges for the ligands with antechamber in the AMBER package. The ff99SB force field was used for the protein with the TIP3P model for the water solvent and the generalised amber force field⁴⁰ for the ligands. The small systems (ligands in water and amino acid pairs) were solvated in such a way that the simulation cells were cubes with faces at least 15 Å away from the atoms of the initial structure of the solute. This resulted in a total of between 1500 and 1600 water molecules in each simulation cell.

E. ONETEP calculations

To generate the ONETEP inputs, the MD trajectories were post processed using the ptraj tool within AMBER. Each system was re-centred so that the solute is at the centre of the

waterbox and the water molecules were indexed according to their distance from the ligand. This facilitated the partitioning of waters to quantum and to classical embedding charges. Snapshots were taken at constant time intervals throughout the trajectory (written as PDB files) to create an ensemble of structures; A few of these structures were randomly selected and used for the tests reported here.

The rms value criterion on the NGWF gradient, as in Ref. 15, was used to determine convergence of the ONETEP calculations, with a NGWF rms gradient threshold of $2 \times 10^{-6} E_h a_0^{3/2}$. However, we have additionally checked that the total energies were converged to at least 0.0002 E_h (~ 0.1 kcal mol⁻¹) even for the largest calculations, such as the lysozyme-phenol complex in Sec. III A 2 surrounded by a 12 Å thick shell of quantum waters, which contained 10151 atoms in total. Given that SCF convergence criteria lead to errors in the energy that are constant per atom but increase with the system size, if one wants to calculate accurate energy differences on even larger systems, the tightness of the gradient convergence criterion would need to be increased. It is conceivable, however, that beyond some system size the subtraction of total energies will no longer be a practical way of calculating interaction energies and an alternative approach will be needed. We did not encounter any convergence problems for the calculations reported here and the number of NGWF SCF iterations needed to converge to our NGWF rms gradient threshold showed very little variation with system type or the presence of embedding charges or not. For example, 15 iterations are needed to converge a configuration of the cysteine zwitterion of Sec. III A 1 when surrounded by its first solvation sphere (22 water molecules) with and without embedding charges, 22 iterations for the same molecule in a simulation cell filled with explicit quantum waters (1518 water molecules) while the 10151-atom lysozyme-phenol-water structure mentioned above required 19 iterations, again regardless of whether embedding charges are included in the calculation or not.

Four NGWFs were used on each heavy atom and one for each hydrogen with NGWF localisation radii of 3.7 Å. A kinetic energy cutoff of 800 eV was used for the psinc basis set and the PBE (Ref. 41) generalised gradient approximation (GGA) exchange-correlation functional was used augmented with dispersion contributions (DFT+D approach).³⁴ In these calculations the density kernel was not truncated, for a number of reasons. As we have shown in previous work,¹⁸ even for systems with a regular spatial distribution of atoms such as crystalline silicon the truncation of the kernel has to be validated carefully and the self-consistent convergence tolerance threshold raised accordingly. Biomolecular assemblies such as proteins or molecules in explicit water possess an irregular spatial distribution of atoms with different levels of electronic “nearsightedness” along different directions. As a result, much larger thresholds for the density kernel truncation than in the case of crystalline silicon would need to be used in order to maintain convergence at the level of 0.1 kcal/mol in the calculations. This, combined with the fact that GGA calculations artificially lower the bandgap so that in biomolecular systems in explicit water it is nearly zero,⁴² means that the systems we study here are still too small for kernel truncation

if this level of precision in the energy is desired. Similar conclusions about the effect of GGA functionals can be reached from the work of Rubensson and Rudberg⁴³ who have investigated the bandgaps and the decay of the density matrix in water clusters and other systems and proposed various matrix truncation schemes. However, we should note that there is a significant difference between such atomic orbital based linear-scaling matrix truncation approaches and our work as in ONETEP the choice to not truncate applies only to the density kernel while the other matrices (such as Hamiltonian and overlap) remain sparse by construction due to the strict localisation of the NGWFs. As a result, when the kernel is not truncated the non-linear computational cost is limited only to matrix operations involving the density kernel. The major computational cost of constructing the overlap and Hamiltonian matrices remains linear-scaling and in practice this allows calculations with several thousand atoms to be performed routinely. We should clarify here that while the construction of the Hamiltonian operator (and consequently the matrix) is linear-scaling, in practice, in our applications to date, asymptotically it is not strictly linear-scaling but $\sim V \ln V$, where V is the volume of the simulation cell which can be proportional to the number of atoms. This happens because the Coulomb potential is obtained via solution of the Poisson equation with FFTs.⁴⁴ The $\ln V$ dependence is too small to be observable in any of the scaling tests we have done so far with up to tens of thousands of atoms, but we expect that future enhancements in the parallel algorithms of the code as well as the availability of more powerful computational platforms will lead to calculations on even larger systems with hundreds of thousands of atoms where eventually the $\ln V$ dependence will become observable. This will then need to be tackled with further developments such as, for example, confining the use of FFTs only to the shorter length scales and employing approaches such as hierarchical multiple expansions for longer-range interactions.

Our calculations were performed on the Iridis3 and HECToR supercomputers which are distributed memory clusters. Depending on system size, each of our calculations used from 24 to 256 cores.

III. RESULTS AND DISCUSSION

A. Interaction energies

The electrostatic embedding approach developed is expected to have numerous uses in terms of including long-range electrostatic interactions between a quantum system and its surroundings. One of our main interests in using this approach is to be able to describe a large portion of the surrounding water in biomolecular systems. Particularly useful in this context is the ability to accurately calculate interaction energies as these can be used to obtain Gibbs free energies of binding which are essential in drug design. A variety of methods are available for this such as the recent approach by Beierlein *et al.*⁴⁵ who have demonstrated via a series of careful tests how the free energies obtained via a classical force field can be converted to free energies that would be obtained if a quantum description was used for the ligand and classical force field for the surrounding atoms. The mutation from the

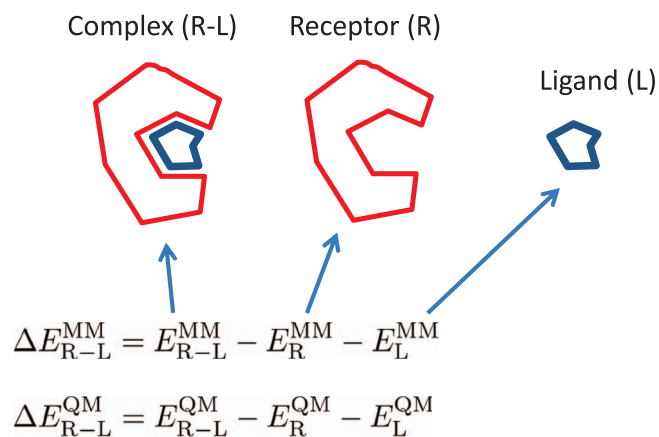


FIG. 3. Receptor-ligand complex. The interaction energies are obtained by single point energy calculations of the complex and of the receptor and ligand in the same geometry as in the complex.

classical to the quantum state happens via a one-step free energy perturbation formula (the Zwanzig equation), as follows:

$$\Delta G_{MM \rightarrow QM} = -k_B T \ln \langle e^{-(E_{R-L}^{QM} - E_{R-L}^{MM})/k_B T} \rangle_{MM}, \quad (15)$$

where E_{R-L}^{QM} is the energy of the receptor-ligand complex in the quantum description and E_{R-L}^{MM} is the energy in the force field description, and the notation $\langle \cdot \cdot \cdot \rangle_{MM}$ signifies an ensemble average over the structures obtained from the MD simulations with the MM force field. While Eq. (15) is formally correct, it is difficult to use in practice due to the large differences in magnitude between the quantum and classical energies which are made up of rather different energy contributions. A more practical form can be obtained if the total energies are substituted by interaction energies,⁴⁵

$$\Delta G_{MM \rightarrow QM} = -k_B T \ln \langle e^{-(\Delta E_{R-L}^{QM} - \Delta E_{R-L}^{MM})/k_B T} \rangle_{MM}, \quad (16)$$

where ΔE_{R-L}^{QM} is the interaction energy of the receptor-ligand complex in the quantum description and ΔE_{R-L}^{MM} is the interaction energy in the force field description, obtained as outlined in Figure 3.

Our goal is to extend such approaches to a quantum mechanical treatment of the ligand and a number of atoms from the surroundings which is large enough to ensure that the interaction energies obtained will be converged to chemical accuracy. While this can be achieved in a purely brute force manner by simply increasing the size of the quantum region, the use of electrostatic embedding should significantly speed up the convergence of interaction energies. The aim is to be able to obtain interaction energies without any appreciable change in their values as compared to the fully quantum result (no embedding, all atoms are quantum). In Secs. III A 1–III A 2, we investigate the effect of embedding in a variety of systems ranging from ligands in water to ligands in an entire protein.

1. Solvent-ligand interactions

Solvent-ligand interactions are important as solvent is encountered in almost all practical applications. We will confine our study to water as it is the most common solvent, espe-

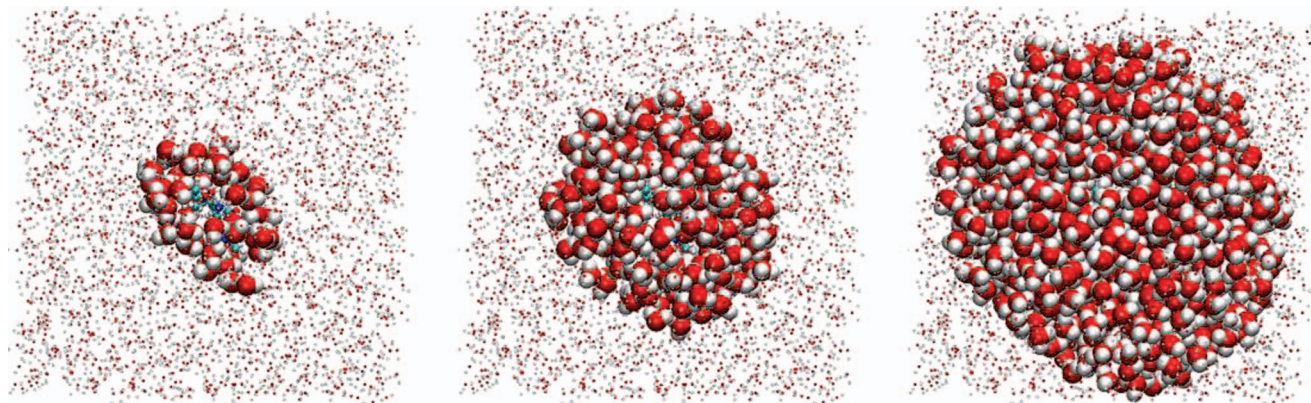


FIG. 4. Separation of the phenol-water system into quantum and embedding atoms. From left to right, 50, 250, and 750 water molecules closest to the ligand are treated as quantum atoms within the ONETEP calculation. The remaining water molecules of the simulation (which was carried out in a waterbox of about 1600 water molecules) are treated as classical embedding charges.

cially in biomolecular systems. As ligands for our tests we have used the molecules toluene, bromobenzene, phenol, thiophenol, catechol (polar), cysteine terminated by N-terminal acetyl (ACE) and C-terminal N-methyl (NME) groups, cysteine zwitterion, and serine zwitterion (a polar amino acid). We have calculated interaction energies for each of these ligands, where the water molecules in the simulation cell play the role of the “receptor,” for two MD snapshots that were randomly selected from the “production” stage of our molecular dynamics simulations. For each such snapshot interaction energies were obtained by including around the ligand a successively increasing number of water molecules from the simulation cell waterbox. The waters were included as “solvation shells” of increasing radius, centred on the ligand, as demonstrated in the example shown in Figure 4. Interaction energies were obtained with three approaches:

1. Molecular mechanics force field (ff99SB) calculations for the ligand and the surrounding water solvation shell (MM).
2. ONETEP DFT calculations for the ligand and the surrounding water solvation shell (QM).
3. ONETEP DFT calculations for the ligand and the surrounding water solvation shell, including the remaining water molecules of the waterbox via electrostatic embedding (QM EE).

The interaction energies as a function of the number of surrounding waters as obtained with the above-mentioned three approaches are shown in Figure 5. The energies for the two snapshots (e.g., MM1 is for snapshot 1 and MM2 for snapshot 2 in the MM calculations) are shown for each ligand. In all cases the MM energies converge to the full waterbox result with increasing number of waters more smoothly and more rapidly than the QM energies. This is to be expected as no charge polarisation or movement is possible in the MM case. The embedded QM (QM EE) calculations afford the smoothest and most rapid convergence from all cases. This indicates that the combination of a certain number of quantum waters around the solute, with electrostatic embedding to represent the water occupying the remaining simulation cell, does capture adequately all the charge polarisation of the lig-

and and the back-polarisation of its surroundings that is characteristic of the quantum description. This approach also describes correctly the long range electrostatic interactions as they emerge from the periodic boundary conditions that apply to both our MD and quantum calculations. A physically meaningful measure for deciding how many quantum waters to include in the calculations can be provided by the radii of the solvation shells. For example, for the case of the phenol molecule in Figure 5 the energies are obtained for shells of water with approximate distances from the atoms of the molecule of about 3.4 Å (first solvation shell), 5 Å (second solvation shell), 9 Å, 12 Å, 14 Å, and 17 Å. For all the examples of small ligands in Figure 5, we can conclude that including shells of quantum water up to 12 Å (which consists of about 400 quantum water molecules) combined with electrostatic embedding is enough to produce interaction energies which are virtually indistinguishable from the full quantum system which contains about 1500 water molecules. Both types of calculation are feasible with a code such as ONETEP but the former requires significantly less computational effort (336 core hours as compared to 2800 core hours for the full quantum system). Our observations about the rate of convergence of interaction energies for the QM approach (i.e., without embedding charges) are consistent with the findings of Bondesson *et al.*⁴⁶ who studied with DFT and Hartree-Fock calculations the behaviour of the interaction energy between small drug molecules and solvation shells of explicit quantum waters of increasing thicknesses of up to 14 Å and 732 water molecules. As their calculations were performed with Gaussian basis sets, they found that there is strong dependence on the quality of the basis set and that it is important to include polarisation functions. Also, for such fixed atomic orbital basis sets the correction for BSSE is necessary and remains so even when larger basis sets are used,⁴⁷ as we also demonstrate in our validation tests in Table I. Their calculations did not however include dispersion interactions, which are a significant contributor to the interaction energy. For example, the dispersion component of the interaction energy for snapshot 2 of the catechol molecule in Figure 5 is -19.1 kcal/mol. Cabral do Couto *et al.*⁴⁸ have also used shells of quantum water molecules surrounded by electro-

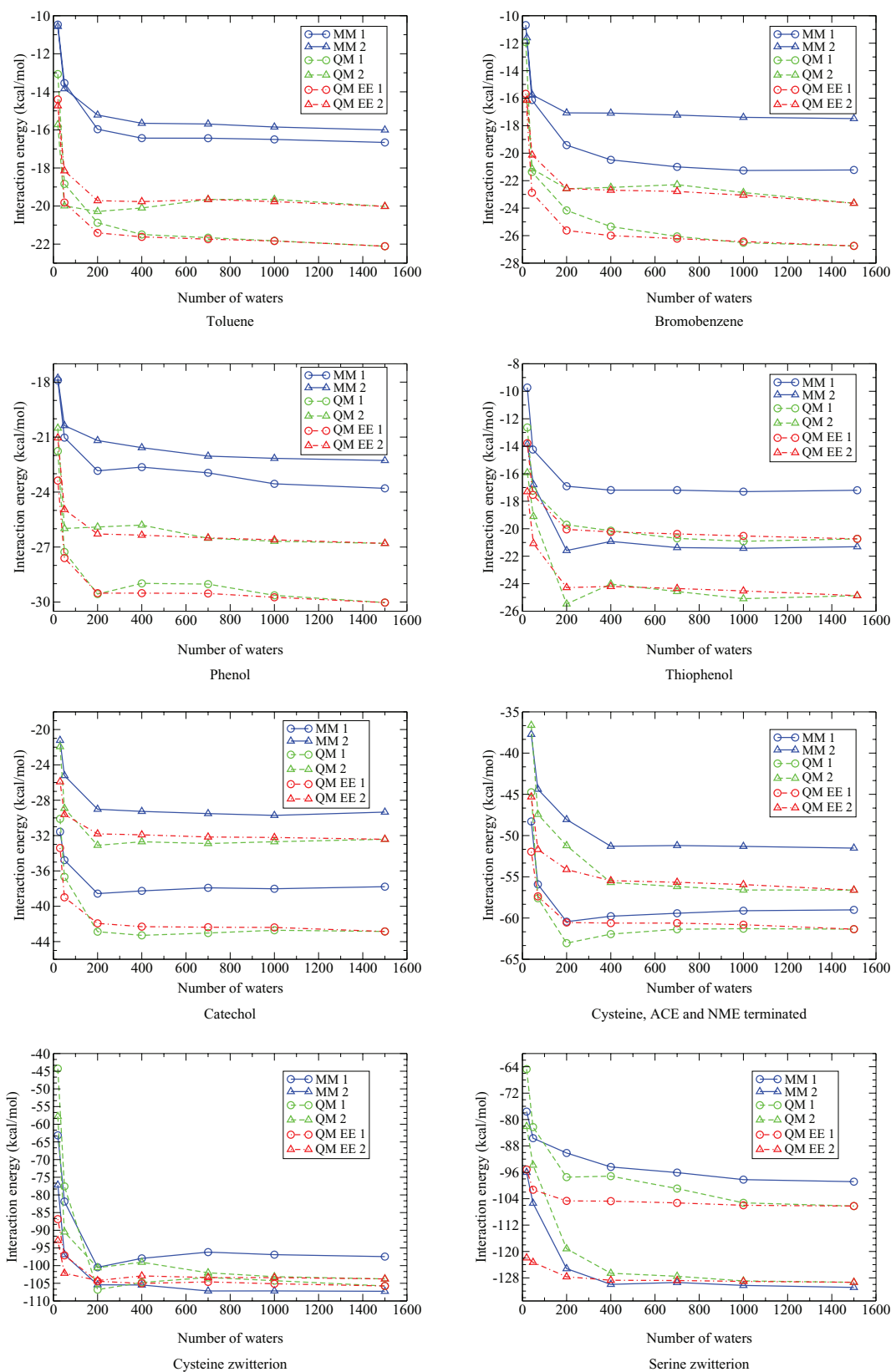


FIG. 5. Interaction energies between ligands and water as a function of increasing number of water molecules.

static embedding charges in order to reduce the undesired surface effects when studying the electronic properties of a water molecule in bulk water.

A further measure of the performance of the embedding approach can be provided by investigation of its effect on

atomic charges, which are indicators of the chemical environment that the atoms experience. In Figure 6, we investigate the Mulliken atomic charges of the atoms of the cysteine zwitterion for increasing sizes of solvation regions for the QM and QM EE approaches. Taking the fully quantum calculation as

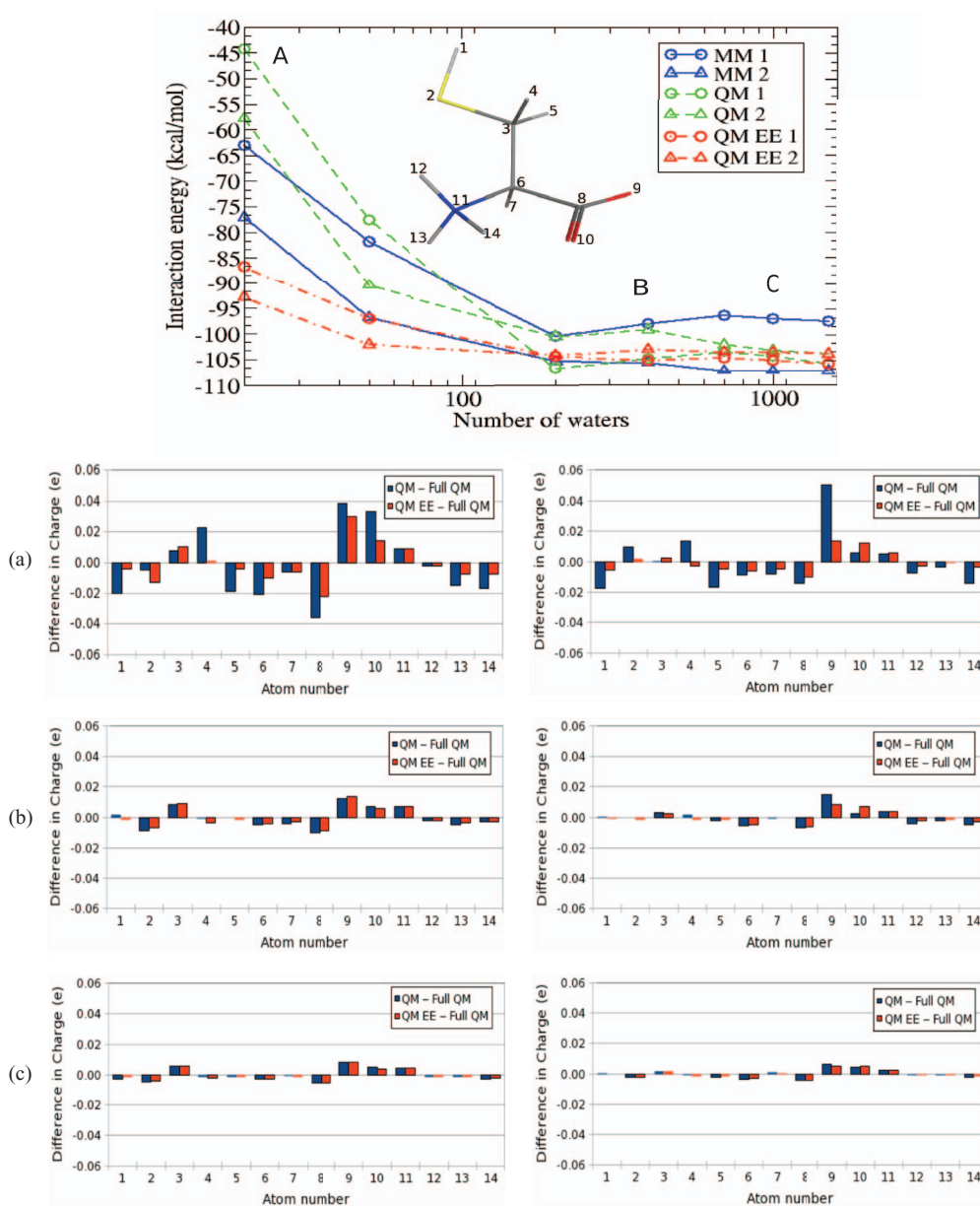


FIG. 6. Variation of atomic charges from the quantum calculation on the cysteine molecule as a function of the thickness of the solvation shell, for snapshot 1 (left) and snapshot 2 (right) for (a) 20 quantum waters, (b) 400 quantum waters, and (c) 1000 quantum waters. For each solvation shell, the difference of the charge on each atom from the charge obtained from the full QM calculation (including all waters in the simulation cell in the quantum description) is given for the QM and QM EE approaches.

the benchmark, we observe that for the first solvation sphere the QM EE approach produces for most atoms less than half the error of the QM calculation. However, for the case of 400 quantum waters or more the differences between QM and QM EE diminish as the errors become small (less than 0.01 eV), and for this case the difference between the QM and QM EE interaction energies is small.

2. Receptor-ligand interactions

We next evaluate the performance of the electrostatic embedding approach on biomolecular assemblies treated in their entirety by quantum calculations. We seek to use electrostatic embedding to include the electrostatic effects of the environ-

ment (usually solvent molecules). For this task we consider a number of receptor-ligand complexes. Initially, we consider small complexes where the receptor is either a lysine (LYS) (an amino acid with net charge of +1) or ASP (an amino acid with net charge of -1) and the ligand is a serine (SER) (an amino acid with net charge 0). The interaction energies obtained for these cases are shown in Figure 7, for increasing numbers of water molecules surrounding the receptor-ligand complex.

As the receptors are charged, electrical neutrality is imposed by the presence of a counterion (Cl^- in the case of LYS and Na^+ in the case of ASP-treated as embedding charges in the ONETEP calculations. The Gaussian smeared embedding charges ensure that electronic charge spilling to the Na^+

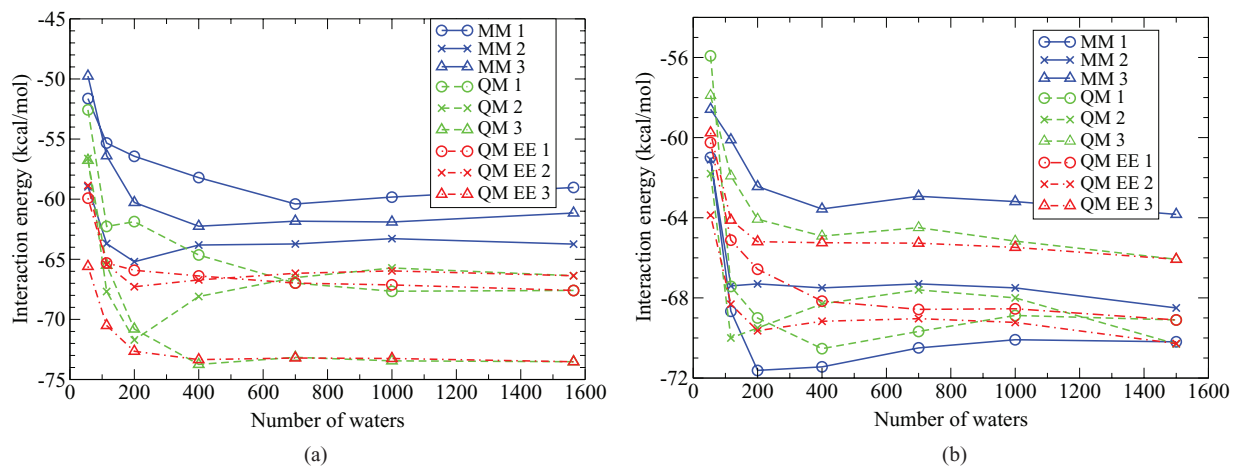


FIG. 7. Interaction energies between a serine (SER) and a lysine (LYS) in water (a) and between serine and an aspartate (ASP) in water (b).

is avoided. We have confirmed that this is indeed the case by plotting and examining isosurfaces of the electronic density). In the plots, we show interaction energies for three snapshots for each receptor-ligand complex. These snapshots were picked at varying distances of the counterion from the receptor-ligand complex, at 10 Å in snapshot 1, to 17 Å in snapshot 2, to 24 Å in snapshot 3 for cubic simulation cells of side 36.5 Å. These variations do not have any observable effect in the rate of convergence of the energies. Interaction energies converge following patterns similar to those of the neutral ligands in water (Figure 5) for the electrostatic embedding approach. The non-embedded calculations converge in a less stable fashion as a result of the presence of the charge separation which is shielded to differing extents by the water layers of increasing thickness, and very slowly: in some cases convergence to the fully quantum result is reached only when the entire simulation cell is filled with water.

To examine the effect of the inclusion of the solvent in a real protein-ligand complex we have considered the T4

lysozyme (L99A/M102Q) protein which is a well-studied model of a polar binding site.⁴⁹ We have examined the complex between T4 lysozyme (L99A/M102Q) and phenol (PDB ID: 1LI2), in water. As the complex has a total charge of +8, electrical neutrality is imposed by including 8 Cl⁻ counterions in the simulation cell, which contains the complex and 9053 water molecules. The interaction energies were calculated as in the scheme of Figure 3, where in this case E_{R-L} is the energy of the entire system (lysozyme, water, and phenol), E_R is the energy of the lysozyme and water and E_L is the energy of the phenol. Figure 8 shows the interaction energies obtained for three snapshots of the complex as well as a picture of one of the snapshots showing the atoms of the protein, the waters that are described as quantum atoms and the waters that are represented as embedding charges. Water is added in spherical shells centered around the ligand, which is located in the binding pocket of the protein. DFT calculations have been performed for zero quantum waters (where only the 2601 protein atoms plus the 13 ligand atoms are described by

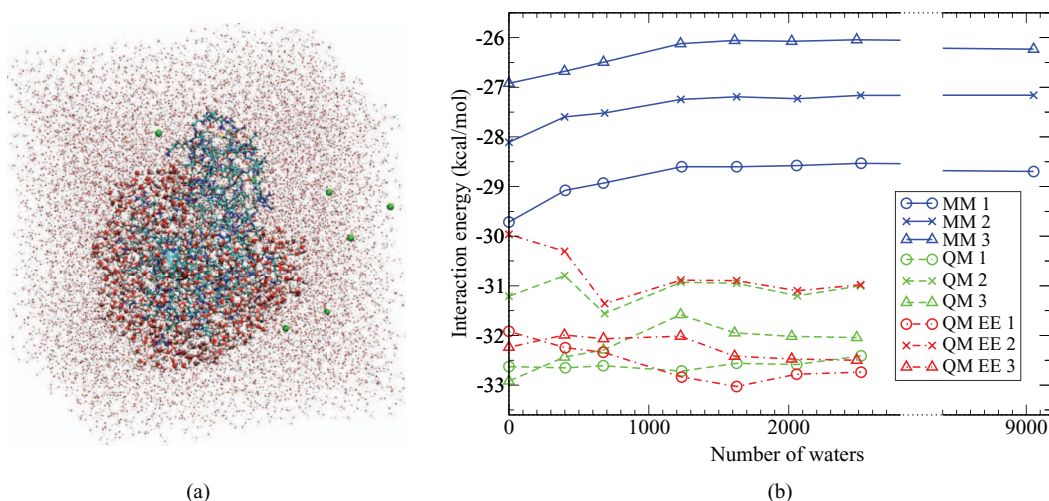


FIG. 8. (a) The complex of L99A/M102Q T4 lysozyme and phenol in water. The second solvation sphere around the ligand binding cavity is shown in ball and stick representation while the rest of the waters are shown as dots. (b) Interaction energies between phenol and L99A/M102Q T4 lysozyme in water for increasing numbers of water molecules within a sphere around the binding pocket.

DFT), and by including shells with thickness of 3.4 Å, 5.0 Å, 7.5 Å, 9.0 Å, 10.5 Å, up to 12.0 Å which results in 10151 atoms in total being treated by DFT.

We can observe for the T4 lysozyme (L99A/M102Q) protein that the presence of waters has a very small effect on the interaction energies. The embedded calculation shows marginally better convergence with respect to the quantum calculation without embedding but the advantage of embedding is almost negligible with variations which are less than 1 kcal/mol. This is a result of the fact that the cavity of T4 lysozyme (L99A/M102Q) is completely buried and shielded from the solvent. In complexes with more exposed cavities the inclusion of water is expected to have a larger effect, depending on the degree of exposure to the solvent. The regions and thickness of quantum water layers that need to be included will vary from one protein to another and need to be determined on a case by case basis. It is interesting however to observe that in the case of this protein the use of electrostatic embedding with no quantum waters leads to the largest errors as the embedding atoms in contact with the quantum atoms of the protein appear to over-polarise it. In fact Figure 8 shows that the DFT calculation with no water at all would be in this case the best compromise between accuracy and efficiency as the errors that result are of the order of 0.5 kcal/mol which is comparable to other errors intrinsic in DFT calculations (such as the choice of exchange-correlation functional and the basis set). Indeed, the error in the interaction energies with respect to the basis set in this case is expected to be of the order of 1.2 kcal/mol from the validation tests of Table I for NGWF radii of 3.7 Å. Another interesting observation is that the QM and QM EE curves, for snapshots 1 and 3, do not coincide, even for the largest calculations with 2517 quantum water molecules. As the total energies in our calculations were converged to 0.1 kcal/mol (Sec. II E) it is unlikely that this is due to numerical noise but most likely it is a manifestation of the long-range nature of electronic polarisation which appears to not be completely converged for these structures even with this large number of water molecules.

IV. CONCLUSIONS

We have implemented and evaluated a scheme for electrostatic embedding of DFT calculations within a set of Gaussian charge distributions and have outlined its implementation within the psinc basis set framework of the linear-scaling DFT code ONETEP. As the psinc basis set is equivalent to a plane wave basis set, the implementation of the scheme as outlined here could be almost carried out in an essentially identical way on any plane wave basis set code. Usually in QM/MM approaches the QM and MM interface is required to “cut” through chemical bonds due to the limitation in the size of the quantum region that is feasible to work with when using conventional cubic-scaling DFT methods. With ONETEP it is possible to perform DFT calculations on entire macromolecular entities such as molecules, proteins and nanostructures. Some of their surrounding environment (such as the solvent) can also be described at the DFT level, and we advocate using electrostatic embedding in order to account for the remaining long-range electrostatic interactions.

We have tested the scheme in the calculation of interaction energies between molecules and the solvent, and between protein-ligand-solvent complexes and ligands. We have shown that this scheme offers results of the same quality as when describing the entire system (solute and explicit solvent) by quantum calculations, but at much lower computational cost. As the computational cost of the embedding atoms is essentially zero compared to that of the quantum atoms, the reduction in the cost of the calculation is due to the reduction in the number of quantum atoms.

Nevertheless, even with electrostatic embedding, in the case of interactions with the solvent, a significant number of solvent molecules which are described by DFT need to be included. For example, we found that quantum water shells of thickness 12–13 Å need to be included for the interaction energies to converge to chemical accuracy. For small ligands, this results in about 400 water molecules described by DFT while for the ASP-SER and LYS-SER complexes this results in about 700 DFT water molecules. In light of this observation, results obtained with previous QM/MM approaches^{11,45,50} which treat only the solute by QM and the solvent by MM would need to be re-examined as for example calculations of free energies and other properties using these approaches may be dominated by the variation (noise) that the interaction energies have when zero or a very small number of quantum waters are included. We expect that one of the first applications of our embedding approach will be in schemes for the calculation of free energies of binding of biomolecular assemblies with full inclusion of charge transfer and polarisation effects via large-scale DFT calculations.

ACKNOWLEDGMENTS

S.F. would like to thank the BBSRC and Boehringer Ingelheim for an industrial CASE studentship. C.P. would like to thank the BBSRC and Accelrys for an industrial CASE studentship. C.-K.S. would like to thank the Royal Society for a University Research Fellowship. The calculations in this work were carried out on the Iridis3 Supercomputer of the University of Southampton and on the HECToR national supercomputer. Access to HECToR was provided via the Engineering and Physical Sciences Research Council (United Kingdom) (EPSRC) Grant No. EP/F038038/1 (UKCP consortium).

¹A. Warshel and M. Levitt, *J. Mol. Biol.* **103**, 227 (1976).

²I. Solt, P. Kulhanek, I. Simon, W. Steven, M. C. Payne, G. Csanyi, and M. Fuxreiter, *J. Phys. Chem. B* **113**, 5728 (2009).

³H. Lin and D. G. Truhlar, *Theor. Chem. Acc.* **117**, 185 (2007).

⁴M. M. Senn and W. Thiel, *Curr. Opin. Chem. Biol.* **11**, 182 (2007).

⁵F. Claeysens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel, and H.-J. Werner, *Angew. Chem., Int. Ed.* **45**, 6856 (2006).

⁶N. Bernstein, J. R. Kermode, and G. Csanyi, *Rep. Prog. Phys.* **72**, 026501 (2009).

⁷J. Dziedzic, M. Bobrowski, and J. Rybicki, *Phys. Rev. B* **83**, 224114 (2011).

⁸L. Pejov, D. Spångberg, and K. Hermansson, *J. Phys. Chem. A* **109**, 5144 (2005).

⁹G. A. Cisneros, J. P. Piquemal, and T. A. Darden, *J. Phys. Chem. B* **110**, 13682 (2006).

¹⁰T. Schwabe, J. M. H. Olsen, K. Sneskov, J. Kongsted, and O. Christiansen, *J. Chem. Theor. Comput.* **7**, 2209 (2011).

- ¹¹K. E. Shaw, C. J. Woods, and A. J. Mulholland, *J. Phys. Chem. Lett.* **1**, 219 (2010).
- ¹²T. A. Wesolowski and A. Warshel, *J. Phys. Chem.* **97**, 8050 (1993).
- ¹³E. Prodan and W. Kohn, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 11635 (2005).
- ¹⁴S. Goedecker, *Rev. Mod. Phys.* **71**, 1085 (1999).
- ¹⁵C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, *J. Chem. Phys.* **122**, 084119 (2005).
- ¹⁶C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez, and M. C. Payne, *Phys. Rev. B* **66**, 035119 (2002).
- ¹⁷A. A. Mostofi, P. D. Haynes, C.-K. Skylaris, and M. C. Payne, *J. Chem. Phys.* **119**, 8842 (2003).
- ¹⁸C.-K. Skylaris and P. D. Haynes, *J. Chem. Phys.* **127**, 164712 (2007).
- ¹⁹P. D. Haynes, C.-K. Skylaris, A. A. Mostofi, and M. C. Payne, *Chem. Phys. Lett.* **422**, 345 (2006).
- ²⁰T. P. Straatsma, E. Apra, T. L. Windus, M. Dupuis, E. J. Bylaska, W. de Jong, S. Hirata, D. M. A. Smith, M. T. Hackler, L. Pollack, R. J. Harrison, J. Nieplocha, V. Tipparaju, M. Krishnan, E. Brown, G. Cisneros, G. I. Fann, H. Fruchtl, J. Garza, K. Hirao, R. Kendall, J. A. Nichols, K. Tsemekhman, M. Valiev, K. Wolinski, J. Anchell, D. Bernholdt, P. Borowski, D. Clark, T. Clerc, H. Dachsel, M. Deegan, K. Dylla, D. Elwood, E. Glendening, M. Gutowski, A. Hess, J. Jaffe, B. Johnson, J. Ju, R. Kobayashi, R. Kutteh, Z. Lin, R. Littlefield, X. Long, B. Meng, T. Nakajima, S. Niu, M. Rosing, G. Sandrone, M. Stave, H. Taylor, G. Thomas, J. van Lenthe, A. Wong, and Z. Zhang, "NWChem, a computational chemistry package for parallel computers, version 5.1.1, Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA," 2008.
- ²¹S. F. Boys and F. Bernardi, *Mol. Phys.* **19**, 553 (1970).
- ²²C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, *Phys. Status Solidi B* **243**, 973 (2006).
- ²³N. D. M. Hine, P. D. Haynes, A. A. Mostofi, C.-K. Skylaris, and M. C. Payne, *Comput. Phys. Commun.* **180**, 1041 (2009).
- ²⁴P. D. Haynes, C.-K. Skylaris, A. A. Mostofi, and M. C. Payne, *J. Phys. Condens. Matter* **20**, 294207 (2008).
- ²⁵N. Zonias, P. Lagoudakis, and C.-K. Skylaris, *J. Phys. Condens. Matter* **22**, 025303 (2010).
- ²⁶L. Heady, M. Fernandez-Serra, S. Joyce, A. R. Venkitaraman, E. Artacho, C.-K. Skylaris, C. Ciacchi, and M. C. Payne, *J. Med. Chem.* **49**, 5141 (2006).
- ²⁷D. J. Cole, C.-K. Skylaris, E. Rajendra, A. R. Venkitaraman, and M. C. Payne, *Europhys. Lett.* **91**, 37004 (2010).
- ²⁸S. Fox, H. G. Wallnoefer, T. Fox, C. S. Tautermann, and C.-K. Skylaris, *J. Chem. Theor. Comput.* **7**, 1102 (2011).
- ²⁹J. T. Berryman, S. E. Radford, and S. A. Harris, *Biophys. J.* **97**, 1 (2009).
- ³⁰M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, *Rev. Mod. Phys.* **64**, 1045 (1992).
- ³¹C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, C. J. Pickard, and M. C. Payne, *Comput. Phys. Commun.* **140**, 315 (2001).
- ³²R. M. Martin, *Electronic Structure. Basic Theory and Practical Methods* (Cambridge University Press, Cambridge, England, 2004).
- ³³M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, New York, 1987).
- ³⁴Q. Hill and C.-K. Skylaris, *Proc. R. Soc. London, Ser. A* **465**, 669 (2009).
- ³⁵S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, *J. Chem. Phys.* **132**, 154104 (2010).
- ³⁶A. Laio, J. VandeVondele, and U. Rothlisberger, *J. Chem. Phys.* **116**, 6941 (2002).
- ³⁷J.-L. Fattebert, R. J. Law, B. Bennion, E. Y. Lau, E. Schwegler, and F. C. Lightstone, *J. Chem. Theor. Comput.* **5**, 2257 (2009).
- ³⁸Chemical Computing Group Inc., *Molecular Operating Environment (MOE), 2010.10*, 1010 Sherbrooke Street West, Suite No. 910, Montreal, QC, Canada, H3A 2R7 (2010).
- ³⁹D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin, and P. A. Kollman, AMBER 10, University of California, San Francisco, 2008.
- ⁴⁰J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, *J. Comput. Chem.* **25**, 1157 (2004).
- ⁴¹J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- ⁴²C. M. Isborn, N. Luehr, I. S. Ufimtsev, and T. J. Martinez, *J. Chem. Theor. Comput.* **7**, 1814 (2011).
- ⁴³E. H. Rubensson and E. Rudberg, *J. Comput. Chem.* **32**, 1411 (2011).
- ⁴⁴N. D. M. Hine, J. Dziedzic, P. D. Haynes, and C.-K. Skylaris, *J. Chem. Phys.* **135**, 204103 (2011).
- ⁴⁵F. R. Beierlein, J. Michel, and J. W. Essex, *J. Phys. Chem. B* **115**, 4911 (2011).
- ⁴⁶L. Bondesson, E. Rudberg, Y. Luo, and P. Salek, *J. Phys. Chem. B* **111**, 10320 (2007).
- ⁴⁷L. Bondesson, E. Rudberg, Y. Luo, and P. Salek, *J. Comput. Chem.* **29**, 1725 (2008).
- ⁴⁸P. Cabral do Couto, R. C. Guedes, and B. J. Costa Cabral, *Braz. J. Phys.* **34**, 42 (2004).
- ⁴⁹B. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet, *J. Mol. Biol.* **322**, 339 (2002).
- ⁵⁰C. J. Woods, F. Manby, and A. Mulholland, *J. Chem. Phys.* **128**, 014109 (2008).