Large-Scale DFT Calculations in Implicit Solvent—A Case Study on the T4 Lysozyme L99A/M102Q Protein

Jacek Dziedzic,[a],† Stephen J. Fox,[a] Thomas Fox,[b] Christofer S. Tautermann,[b] and Chris-Kriton Skylaris*[a]

Recently, variants of implicit solvation models for first principles electronic structure calculations based on a direct solution of the nonhomogeneous Poisson equation in real space have been developed. These implicit solvation models are very elegant from a physical point of view as the solute cavity is defined directly via isosurfaces of the electronic density, and the molecular charge is polarized self-consistently by the reaction field of the dielectric continuum which surrounds the solute. Nevertheless, the implementation of these models is technically complex and requires great care. A certain level of care is required from users of such models as a number of numerical parameters need to be given appropriate values to obtain the most accurate and physically relevant results. Here, we describe in what parts of the solvent model each of these numerical parameters is involved and present a detailed study of how they can affect the calculation, using the solvation model which has been implemented in the ONETEP program for linear-scaling density functional theory (DFT) calculations. As ONETEP is capable of DFT calculations with thousands of atoms, we focus our investigation of the numerical parameters with a case study on protein–ligand complexes of the entire 2602-atom T4 Lysozyme L99/M102Q protein. We examine effects on solvation energies and binding energies, which are critical quantities for computational drug optimization and other types of biomolecular simulations. We propose optimal choices of these parameters suitable for routine "production" calculations. © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/qua.24075

Introduction

Chemistry, biochemistry, and materials and interfacial processes typically take place in and require the presence of solvent. Therefore, simulations at the atomic level must include a description of the solvent. Implicit solvent models, which describe the solvent as a dielectric continuum, have proved very effective in this task and have been an active area of research with many improvements over the years, both within atomistic classical force field simulation methods and in first principles quantum chemistry methods. These models are particularly effective in the context of quantum chemistry calculations, as the reaction field of the dielectric is included directly in the Hamiltonian operator and polarizes the density during the self-consistent solution of the quantum mechanical model. Notable variants of such self-consistent implicit solvation models are the polarizable continuum model (PCM) of Tomasi and coworkers,[1] the COSMO model[2] as well as the very accurate but heavily parameterized SMD model of Truhlar and coworkers,[3] which is founded in the integral equation formalism[4] of the PCM model. Although the physical principles on which these models are based are very elegant, the actual implementation can depend on a large number of parameters which need careful determination by fitting to experimental or theoretical data.

Recently, Fattebert and Gygi[5] proposed a new model of continuum solvation, where the dielectric is defined as a functional of the electronic density of the solute. This model was further extended by Scherlis *et al.*[6] to include the calculation of the cavitation energy, by defining it in terms of the quantum surface of the solute. This model is particularly attractive, as it retains the elegance of the implicit solvent philosophy, as the reaction field is obtained by direct solution of the nonhomogeneous Poisson equation (NPE) in real space:

$$\nabla \cdot (\varepsilon[\rho]\nabla\phi_{\mathrm{NPE}}(\mathbf{r})) = -4\pi\rho_{\mathrm{tot}}(\mathbf{r}), \quad (1)$$

where $\rho(\mathbf{r})$ is the electronic density and $\rho_{\mathrm{tot}}(\mathbf{r})$ is the total density due to electrons and nuclei (or ionic cores in the case of pseudopotentials). Despite this, results obtained with this model in its original formulation were reasonable but significantly less accurate than the conventional approaches such as PCM, especially for charged molecules. We have recently shown[7] how this limitation can be overcome using appropriate boundary conditions, including dispersion interactions with the solvent and redetermining appropriately the two parameters in the functional $\varepsilon[\rho]$. The solvent model by Dziedzic *et al.* has been validated on an extensive set of more than 130 molecules (a representative selection of 20 neutral, 20 cationic, and 20 anionic molecules from Ref. [8], and 71 larger neutral molecules from Refs. [9, 10]) and produces solvation energies that agree with experimental

[a] *J. Dziedzic, S. J. Fox, C.-K. Skylaris*
School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom
E-mail: c.skylaris@soton.ac.uk

[b] *T. Fox, C. S. Tautermann*
Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany

† Also at Faculty of Technical Physics and Applied Mathematics, Gdansk University of Technology, Poland

measurements, with the degree of agreement comparable to that of the SMD approach. Our solvation model has been implemented in the ONETEP[11] program for linear-scaling density functional theory (DFT)[12–14] calculations, which, owing to its linear-scaling algorithms, has the capability of performing very large DFT calculations with many thousands of atoms.[15] This combination of solvation and linear-scaling DFT opens up new possibilities for realistic large-scale simulations of entire biomolecular assemblies or nanostructures in the presence of solvent.

In this article, we describe the main methodological and computational developments on which the electrostatic component of our solvent model is based and test their numerical behavior to provide users of the model with a set of reference data that will be valuable as a guide for the correct application of the model. For details on its nonelectrostatic components, we refer the reader to Ref. [7]. Given that the ONETEP code is intended for large-scale calculations, we have chosen to perform our tests on two complexes of a ligand with an entire protein (T4 Lysozyme) which contains 2602 atoms. We used phenol and toluene as ligands. In "Smeared Core Charges, Boundary Conditions and Defect Correction" section, we describe important components of the solvent model such as the application of open boundary conditions, the smearing of the ionic charges and the defect correction procedure. In "Calculation Details" section, we describe how the protein system was prepared for our simulations and the set up of the calculations within ONETEP. In "Results and Discussion" section, we present extensive benchmark calculations examining the behavior of the different components of the solvent model and their numerical stability. Finally, in Conclusions section, we summarize our findings and suggest the most stable numerical settings for our solvation model.

Smeared Core Charges, Boundary Conditions and Defect Correction

The solution of the NPE [Eq. (1)] is achieved in real space via a multigrid approach.[16,17] Even though the atomic cores are represented by pseudopotentials, in $\rho_{\text{tot}}(\mathbf{r})$ the atomic cores are replaced by Gaussian charge distributions for numerical convenience. This does not alter the simulated physical system, as with the procedure we outline in Appendix A the obtained total electrostatic energy is that due to the pseudopotential cores and not the Gaussians. By representing the atomic cores as smeared charge distributions rather than point charges, it becomes possible to efficiently and accurately solve the Poisson equation for the total electrostatic potential, due to the total charge density. This avoids the singularities associated with point charges, which are especially problematic in the context of multigrid calculations and treats the ionic and electronic charge distributions on equal footing.[17] Gaussian smearing is commonly used,[6,7,17] although cube-shaped charge distributions have been studied[17] as well. Our model uses Gaussian smearing, and we describe the formalism in more detail in Appendix A.

In vacuum, we solve the homogeneous Poisson equation (HPE):

$$\nabla^2 \phi_{\text{HPE}}(\mathbf{r}) = -4\pi\rho_{\text{tot}}(\mathbf{r}) \tag{2}$$

in the simulation cell, Ω with open boundary conditions, that is, we set up Dirichlet boundary conditions of the form

$$V_{\text{BC}}^{\text{vac}}(\mathbf{r}) = \int_{\Omega} \frac{\rho_{\text{tot}}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}\mathbf{r}' \;\text{ for }\; \mathbf{r} \in \partial\Omega \tag{3}$$

on the faces of the simulation cell, $\partial\Omega$. A direct application of Eq. (3), where the integral is replaced with a sum over a Cartesian grid, is impractical. For a particular grid fineness, the associated computational effort scales as $O(L^2V)$, which, for localized charge, implies $O(L^2N)$ (where L^2 represents the area of a face of the simulation cell, V its volume, and N is the number of atoms), and the prefactor is prohibitively large. To reduce the computational cost, a coarse-grained representation $\rho_{\text{tot}}^{\text{CG}}(\mathbf{r})$ of $\rho_{\text{tot}}(\mathbf{r})$ can be used instead. In this work, $\rho_{\text{tot}}^{\text{CG}}(\mathbf{r})$ is constructed as a set of N_{CG} point charges, each of which corresponds to a cubic block of the simulation cell, encompassing $p \times p \times p$ grid points. The magnitude of each point charge is the sum of the charges on the grid points belonging to the block, and the charge is positioned at the center of charge of the block, $\mathbf{R}_I$, which, in general, does not lie on a grid point. Althogh this approach does not help with the unfavorable scaling, it easily reduces the prefactor by 2–3 orders of magnitude simply by replacing the integral in Eq. (3) with a sum over a small number of point charges:

$$V_{\text{BC}}^{\text{vac}}(\mathbf{r}) \approx \sum_{I}^{N_{\text{CG}}} \frac{\rho_{\text{tot}}^{\text{CG}}(\mathbf{R}_I)}{|\mathbf{r}-\mathbf{R}_I|} \quad \text{for } \mathbf{r} \in \partial\Omega. \tag{4}$$

The parameter p can be used to balance accuracy (which increases as p is made smaller) and computational efficiency (as $N_{\text{CG}} \sim p^{-3}$).

In solution, where the NPE (1) needs to be solved, the open boundary conditions can be no longer obtained from Eq. (3). In our approach, we use an approximation, where for the purpose of calculating the boundary conditions, we assume the dielectric permittivity to be homogeneous and to have the bulk value ε_∞ everywhere, that is, we approximate the potential on the faces of the cell as

$$V_{\text{BC}}^{\text{sol}}(\mathbf{r}) \approx \frac{1}{\varepsilon_\infty} \sum_{I}^{N_{\text{CG}}} \frac{\rho_{\text{tot}}^{\text{CG}}(\mathbf{R}_I)}{|\mathbf{r}-\mathbf{R}_I|} \quad \text{for } \mathbf{r} \in \partial\Omega. \tag{5}$$

As in the presence of the solvent the solute molecule is screened by the dielectric which polarizes in response to the charge density of the solute, the values of the potential on the boundaries of the cell are much smaller than in vacuum.

The multigrid solver uses a second-order discretization of the Laplacian operator which, even at numerical convergence, limits the accuracy of the solution to that of a second-order representation on each grid point. It is possible to obtain more accurate higher order solutions by applying, after the second-order solution has converged, an iterative improvement technique known as defect correction,[18,19] which is outlined in Appendix B. With this approach, the order of the defect correction applied is the resulting polynomial order with which the solution of the HPE or NPE is obtained at each grid point. Of course, as the grid spacing becomes finer, it is expected that the second-order solution will asymptotically reach the higher order solution.

Calculation Details

T4 lysozyme L99A/M102Q

Lysozymes are enzymes that act as a natural form of protection from pathogens, forming part of the innate immune system. They destroy bacteria by attacking the carbohydrate chains which are the main component of the bacterial cell wall ("skin") that braces their delicate membrane against the cell's high osmotic pressure. The process involves catalyzing hydrolysis of 1,4-beta-linkages between *N*-acetylmuramic acid and *N*-acetyl-d-glucosamine residues. The lysozyme binds to the bacterial cell wall and destroys its structural integrity so that the bacteria burst under their own internal pressure.

There has been a great deal of research into protein stability, folding, and design by looking at mutations of the lysozyme from the bacteriophage T4.[20–26] T4 lysozyme can only hydrolyze substrates which have peptide side chains bonded to the polysaccharide backbone. One of the many mutants of T4 lysozyme is Leu99Ala/Met102Gln (or L99A/M102Q). This mutation creates a small buried polar cavity which is capable of encapsulating small aromatic ligands. The T4 lysozyme L99A/M102Q has been used to compare and validate binding free energy methods and to develop docking procedures.[21,25] As it is a (relatively) small, stable protein, with wide availability of experimental as well as computer simulation data, it is a good choice for our benchmark calculations.

Preparation of protein coordinates

The X-ray crystal structure of the complex of T4 Lysozyme L99A/M102Q with phenol (PDB: 1LIE) was protonated (except its single histidine, which is solvent accessible, which was protonated manually) using the Protonate3D approach[27] within the MOE program,[28] before solvating with explicit water in a rectangular box with periodic boundary conditions in the AMBER version 10 package.[29] The total charge of the complex was +9 and it was neutralized by including 9 Cl^- anions in the simulation box. The complex consisted of 2615 atoms. Before a "production" molecular dynamics (MD) simulation, an equilibration stage is required. To achieve this within the limited timescales 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 used: the hydrogens were relaxed, all heavy atoms were kept 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 over 200 ps, with the protein still restrained, in the NVT ensemble and ran for a further 200 ps in the NPT ensemble at 300 K. The system was subsequently cooled to 100 K over 100 ps and a series of minimizations was carried out with the restraints on the protein heavy atoms reduced in stages (500, 100, 50, 20, 10, 5, 2, 1, 0.5 kcal mol^{-1} A^{-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 stabilized and so was the internal structure of the protein, as measured by the root mean square deviation of the backbone atoms from the starting structure, which was 0.75 Å. The final coordinates from this equilibration stage were used as starting coordinates for a production simulation for 20 ns with the NVT ensemble at 300 K.

For our MD simulations, we used the Langevin thermostat, the particle mesh Ewald method for the long range electrostatics, a time-step of 2 fs, and the SHAKE algorithm[30] to constrain hydrogen-containing bonds. The AM1-BCC method was used to obtain partial charges for the ligands with antechamber in the AMBER package. The ff99SB forcefield[31] was used for the protein with the TIP3P model[32] for the water solvent and the generalized amber forcefield (gaff)[33] for the ligands.

The complex of T4 lysozyme with toluene was also simulated. For this, we started from the final structure from the equilibration stage of the complex with phenol. In this structure, phenol was replaced by toluene, and a production simulation was run for 20 ns, again with the NVT ensemble at 300 K.

Next, we performed Molecular Mechanic Poisson-Boltzmann Surface Area (MM-PBSA)[34,35] binding free energy calculations on 1000 snapshots from each MD simulation. In each case, we selected the single snapshot that provided the median value of the free energy. The structures for the complex from these snapshots, with explicit water and counterions removed, were used for this study. In each case, the "host" and "ligand" coordinates were extracted from the complex geometry.

Details of the DFT calculations

The ONETEP[11] program is a linear-scaling DFT code that is capable of achieving large basis set accuracy comparable to that of conventional cubic-scaling plane-wave or Gaussian basis set DFT methods. Its novel and highly efficient algorithms allow DFT calculations with tens of thousands of atoms. It is based on a linear-scaling reformulation of DFT in terms of the one-particle density matrix. The density matrix is represented in terms of strictly localized nonorthogonal generalized Wannier functions (NGWFs),[36] $\phi_\alpha(\mathbf{r})$, and the density kernel, **K**, which is the matrix representation of the density matrix in the duals of the NGWFs. Linear-scaling is achieved by truncation of the density matrix, and by enforcing strict localization of the NGWFs onto atomic regions. During calculations with ONETEP both the density kernel and the NGWFs are optimized self-consistently. Kohn–Sham orbitals[13] are not computed at any stage of the calculation and suitable sparse matrix algebra algorithms are used to ensure computational effort that increases linearly with the number of atoms. Optimizing the NGWFs *in situ* allows for a minimum number of NGWFs to be used, while still achieving large basis set accuracy. The NGWFs are expanded in a basis set of periodic sinc (psinc) functions[37] which are equivalent to a plane-wave basis set. Using a plane-wave basis set allows the accuracy to be improved by changing a single parameter, equivalent to the energy cutoff in conventional plane wave DFT codes. The psinc basis set provides a uniform description of space, meaning that ONETEP does not suffer from basis set superposition error.[38]

For the structures described above, single-point energy calculations were performed with ONETEP. A kinetic energy cutoff of 827 eV was used, corresponding to a grid spacing of 0.5 a_0. Charge

Table 1. Total energy (in kcal/mol) in vacuum and in solution of the complex, host, and ligand (phenol: top, toluene: bottom), their corresponding free energies of solvation and the binding energy of the ligand, as a function of the localization radius of the NGWFs (in atomic units).

NGWF Radius	Total energy in vacuum			Total energy in solvent			Free energy of solvation			Ligand binding energy	
	Complex	Host	Ligand	Complex	Host	Ligand	Complex	Host	Ligand	In vacuum	In solution
7	−7360591.8	−7326671.2	−33889.413	−7363006.4	−7329086.5	−33893.387	−2414.56	−2415.32	−3.97	−31.25	−26.51
8	−7361235.1	−7327312.7	−33892.766	−7363645.4	−7329723.3	−33896.764	−2410.33	−2410.55	−4.00	−29.59	−25.37
9	−7361521.6	−7327598.4	−33894.032	−7363938.5	−7330015.7	−33898.112	−2416.93	−2417.33	−4.08	−29.17	−24.69
10	−7361677.9	−7327754.2	−33894.640	−7364102.0	−7330178.4	−33898.799	−2424.08	−2424.21	−4.16	−29.05	−24.77
7	−7354735.8	−7326526.9	−28183.835	−7357210.5	−7329000.4	−28182.489	−2474.73	−2473.48	1.35	−25.06	−27.66
8	−7355377.5	−7327166.7	−28187.388	−7357848.3	−7329636.3	−28185.950	−2470.80	−2469.59	1.44	−23.38	−26.03
9	−7355664.2	−7327452.8	−28188.646	−7358141.8	−7329929.1	−28187.224	−2477.61	−2476.39	1.42	−22.84	−25.47
10	−7355821.4	−7327609.4	−28189.173	−7358306.5	−7330093.3	−28187.790	−2485.03	−2483.86	1.38	−22.83	−25.39

densities were represented on a grid twice as fine, this was also the grid used in the multigrid calculations. Exchange–correlation was described with the PBE functional. Dispersion interactions between the solute atoms were taken into account with a DFT+D approach due to Hill *et al.*,[39] whereas those between the solute and the solvent were approximately modeled in the implicit solvent approach.

To ensure maximum cancellation of errors, the calculations in vacuum used smeared ions and used the multigrid approach in the electrostatics calculations, similarly to the calculations performed in solution. The converged electronic density obtained in vacuum was used to generate the density-dependent dielectric cavity in solution. We have shown earlier[7] that the error incurred by keeping the dielectric fixed (rather than allowing it to respond to the changes in the electronic density throughout the SCF procedure) is modest, and we accordingly kept the dielectric fixed in all calculations reported here.

We have investigated several parameters that are of interest in DFT studies of solvation and binding energies with implicit solvent. First, we verified how the obtained energies converge with the choice of the localization radius of the basis functions (here, NGWFs). To this effect, we performed full calculations in vacuum and in solution, for four different localization radii, for all the structures. Second, we set out to verify the effect of including solute-solute dispersion energy in the calculation. As in the DFT+D approach, this energy term does not depend on the electronic degrees of freedom, the situation with no dispersion could be easily modeled by an *a posteriori* subtraction of this term from the energies. Third, we studied the effect of a parameter of the smeared ion formalism, the ion smearing width. Here too, full calculations in vacuum and in solution were performed, for six smearing widths. This was only done for the systems involving the phenol ligand.

The effect of the remaining two parameters was studied in a simplified manner, to reduce the computational effort. The simplification consisted in performing only single evaluations of the Coulombic energy (A6) for calculations in vacuum and in solution that had previously been converged with the reference parameters. Only the changes in this energy term were considered. The parameters studied in this fashion were: the block size used for charge coarse-graining when determining the boundary conditions (denoted with p in "Smeared Core Charges, Boundary Conditions and Defect Correction" section), and the order of the finite differences used during defect correction. This was also done only for systems involving the phenol ligand.

Calculations of binding energies were performed on the single snapshots of the T4 Lysozyme L99A/M102Q complexes which were produced as described in "Preparation of Protein Coordinates" subsection. The binding energy was obtained as the difference of the energy of the host and ligand from the energy of the complex. The host and ligand geometries were the same as in the complex. No entropic effects were included in the calculations.

Results and Discussion

NGWF radius

In this section, we investigate the effect of the localization radius of the NGWFs on the obtained total energies, free energies of solvation, and binding energies. Accuracy but also computational effort increase with NGWF radius, so it is important to select the smallest NGWF radii which give the required level of accuracy.[40]

We found that the total energies, which we report in Table 1, are well-converged for all the systems, both in vacuum and in solution. The difference between the results obtained with the smallest cutoff of $7\,a_0$ and with the largest $(10\,a_0)$ is less than 0.02% in all cases, whereas assuming a localization radius of $8\,a_0$ (what we propose as a default) leads to differences of about 0.006%, compared with results obtained with a radius of $10\,a_0$. Although the total energies converge monotonically with an increase of the NGWF radius (as it is a variational parameter[41]), their respective differences (i.e., solvation energies) do not. However, for the complex and the host the solvation energies do not vary by more than 0.5% as the NGWF radius is changed. In the case of the ligands, the relative difference is somewhat larger (up to 2%), yet this corresponds to absolute differences of less than 0.2 kcal/mol. Such differences are not unexpected, given that in our implicit solvent model the dielectric permittivity is a function of the electronic density—thus changes in the localization radius of the NGWFs, which slightly affect the electronic density, will in turn lead to changes in the dielectric permittivity and, consequently, solvation energies. Another effect at play here is the inherent incompleteness of the SCF convergence procedure. We verified this by repeating a subset of the calculations with stricter energy convergence thresholds. The observed changes in the obtained

solvation energies of the complex and the host were in the order of 2–3 kcal/mol. This is an inevitable consequence of the fact that the solvation energies are obtained as a difference of two very large quantities (total energies, which in the case of the complex and the host are as big as 7×10^6 kcal/mol). This demonstrates the importance of converging the total energies very accurately.

The binding energies, which are the most important of the energies presented as far as chemical applications are concerned, both in vacuum and in solvent and for both ligands, quickly converge as the NGWF radius is increased, as we demonstrate in Table 1 and in Figure 1. They are already converged to better than 1 kcal/mol ("chemical accuracy") with an NGWF localization radius of 8.0 a_0, for both ligands.

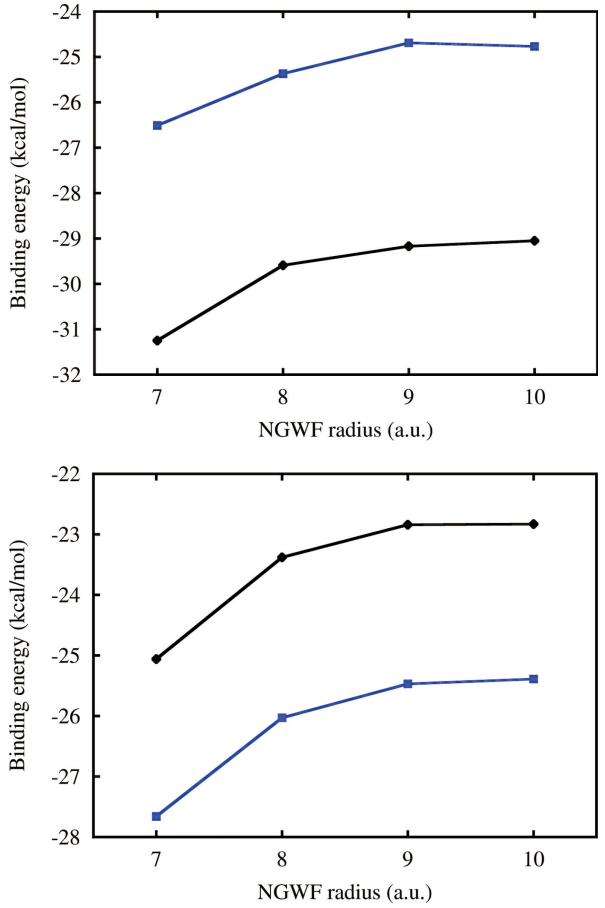

Figure 1. Binding energy (in kcal/mol) of the ligand (phenol: top, toluene: bottom) in vacuum (black diamonds) and in solution (blue squares) as a function of the localization radius of the NGWFs (in atomic units). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

It is interesting to note that absolute binding energies obtained with implicit solvent models are not directly comparable with experimental free energies of binding. For example, the energies of binding for phenol and toluene for NGWF radii of 8.0 a_0 in Table 1 are −25.4 and −26.0 kcal/mol, whereas the experimentally determined free energies of binding are −5.5 and −5.2 kcal/mol[25] respectively. For comparison, the values obtained with the force field described in "Preparation of Protein Coordinates" section with the same Poisson–Boltzmann implicit solvent model as in the MM-PBSA calculations and same coordinates as in Table 1, are −11.7 and −13.8 kcal/mol, respectively. Averaging over many configurations as in the MM-PBSA approach will lead to converged values but not to agreement with experiment as the entropy of binding is not included in such calculations. These entropic terms, however, are expected to cancel out to a great extent when relative free energies of binding with respect to the same protein are considered. In our example, the relative energies of binding are 0.6 kcal/mol for ONETEP, 2.1 kcal/mol for AMBER, and −0.3 kcal/mol for experiment.

Dispersion

Here, we briefly demonstrate the importance of dispersion interactions in simulations of biomolecular association. In particular, it is well known that common generalised gradient approximation (GGA) exchange-correlation functionals fail to correctly describe the attractive component of dispersion interactions. If this deficiency is not corrected (either with DFT+D as we do here,[39] or using nonlocal exchange-correlation functionals that include dispersion[42]), extremely inaccurate energies of binding, both in vacuum and in solvent are obtained. This is apparent from the last two columns of Table 2. This confirms previous reports[39,43] indicating that the inclusion of dispersion in the context of biochemical simulations is essential. We note that in the implicit solvent approach the solute–solvent component of dispersion is modeled implicitly and thus, the values for the free energy of solvation reported in Table 2 are independent of whether solute–solute dispersion has been taken into account.

Smeared ions

The only parameter of the Gaussian smeared ion formalism is the smearing width, σ_I of the Gaussian distributions representing the ions. Although, in principle, this quantity can depend on the species of atom I, for the sake of simplicity, we will assume

Table 2. Total energy (in kcal/mol) in vacuum and in solution of the complex, host, and ligand (phenol: top, toluene: bottom), their corresponding free energies of solvation and the binding energy of the ligand, depending on whether or not dispersion was taken into account.

	Total energy in vacuum			Total energy in solvent			Free energy of solvation			Ligand binding energy	
Dispersion	Complex	Host	Ligand	Complex	Host	Ligand	Complex	Host	Ligand	In vacuum	In solution
yes	−7361235.1	−7327312.7	−33892.766	−7363645.4	−7329723.3	−33896.764	−2410.33	−2410.55	−4.00	−29.59	−25.37
no	−7359163.9	−7325267.0	−33890.826	−7361574.2	−7327677.5	−33894.825	−2410.33	−2410.55	−4.00	−6.06	−1.84
yes	−7355377.5	−7327166.7	−28187.388	−7357848.3	−7329636.3	−28185.950	−2470.80	−2469.59	1.44	−23.38	−26.03
no	−7353314.6	−7325132.4	−28184.631	−7355785.4	−7327602.0	−28183.193	−2470.80	−2469.59	1.44	2.47	−0.18

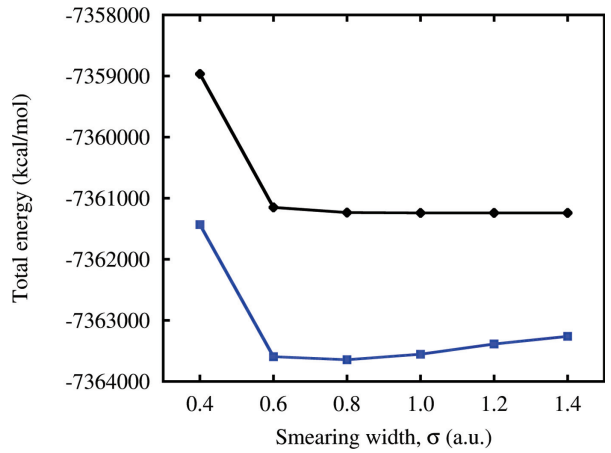

Figure 2. Total energy of the complex (black diamonds—in vacuum, blue squares—in solution) as a function of the smearing width used to smear the ions. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

that the smearing width is identical for all atoms in the system. This section is devoted to the analysis of the influence of this parameter on the obtained total energies of the complex, host, and phenol ligand, their free energies of solvation and the binding energy of the phenol ligand.

Figure 2 shows the total energy of the complex, in vacuum and in solution as the smearing width is varied. These energies are also listed in the first columns of Table 3, along with similar values for the host and the phenol ligand. As detailed in Appendix A, the introduction of smeared ions in vacuum should not, in principle, have any effect on the energy, as both the self-interaction and the nonself-interaction of the Gaussian charge distributions are accounted for in the energy. In practice, however, these charge distributions are represented on a grid with a finite spacing, which inevitably introduces a numerical discretization error which is a consequence of the inability of the grid to accurately represent the Gaussian function as the smearing width is made smaller and becomes comparable with the grid spacing. Here, the grid spacing was $0.25\,a_0$ and a moderate inaccuracy in the total energy can already be observed for $\sigma = 0.6\,a_0$. The inaccuracy becomes apparent for $\sigma = 0.4\,a_0$, where it reaches about 2275 kcal/mol (or 0.03%) for the complex. Similar behavior is observed for the host and the ligand (cf. Table 3). This may give the impression that the problem can easily be alleviated simply by increasing the smearing width, but such conclusion would only be valid in vacuum.

As the dielectric in our implicit model is polarized by the total charge density of the solute, it is essential to ensure that the charge of the smeared ions is well-screened by the electronic charge density. Otherwise, if the ions are smeared too wide, their Gaussian tails approach too close to the cavity boundary, leading to an unphysical depolarization of the dielectric. The fact that even moderate smearing widths already begin to display this effect is shown in Figure 3, which presents cross-sections through charge densities of a hydrogen and a carbon atom, respectively. These plots serve to demonstrate the need to use the smearing width to balance the discretization error (where the thinnest, tallest Gaussians are poorly represented on the grid) with the unphysical behavior arising from excessively broad smearing. Indeed, it is apparent from Figure 2 that the computed energy in solution progressively starts to suffer from the latter inaccuracy as the smearing becomes excessive. For our calculations, we use $\sigma = 0.8\,a_0$ which represents a reasonable compromise. Calculations using finer grids could very well use smaller values of σ, which we demonstrate on the example of the phenol ligand in Figure 4. With a small molecule like phenol, we were able to use a finer grid (with a spacing of $0.125\,a_0$ rather than $0.25\,a_0$). It is clear that with progressively finer grids, it becomes possible to accurately represent Gaussians with smaller smearing widths.

The fact that the total energies in vacuum are not sensitive to large values of the smearing width σ, whereas the energies in solution are, has as a consequence that the free energy of solvation significantly depends on the smearing width. Figure 5 shows this dependence for the complex and for the phenol ligand, serving as evidence that the description of the solvation effect becomes inaccurate if unphysically broad smearings are used. Here again, owing to its smaller size, we can investigate the phenol ligand in more detail—a plot of the its free energy of solvation is shown in Figure 6. Similarly as with total energies, the use of a finer grid allows representing thinner Gaussians accurately, however, it is apparent that even with a very fine grid (with a spacing of $0.125\,a_0$), the use of Gaussians with $\sigma = 0.2$ already leads to numerical inaccuracies. Indeed, the half-width of such a Gaussian is merely $0.167\,a_0$, which is comparable with the grid spacing.

We should note that, to a certain extent, the smearing width effect can be ameliorated using an identical or similar smearing width when parameterizing the solvation model. For instance,

Table 3. Total energy (in kcal/mol) in vacuum and in solution of the complex, host, and phenol ligand, their corresponding free energy of solvation and the binding energy of the ligand in vacuum and in solution, as a function of the smearing width σ of smeared ions (in atomic units).

	Total energy in vacuum			Total energy in solvent			Free energy of solvation			Ligand binding energy	
σ	Complex	Host	Ligand	Complex	Host	Ligand	Complex	Host	Ligand	In vacuum	In solution
0.4	−7358965.5	−7325053.8	−33882.101	−7361432.3	−7327520.7	−33886.429	−2466.8	−2466.9	−4.33	−29.584	−25.188
0.6	−7361150.5	−7327228.5	−33892.370	−7363596.1	−7329674.3	−33896.575	−2445.6	−2445.8	−4.21	−29.611	−25.218
0.8	−7361235.1	−7327312.7	−33892.766	−7363645.4	−7329723.3	−33896.764	−2410.3	−2410.6	−4.00	−29.585	−25.369
1.0	−7361241.0	−7327318.7	−33892.795	−7363555.5	−7329633.6	−33896.136	−2314.4	−2314.9	−3.34	−29.587	−25.789
1.2	−7361241.5	−7327319.0	−33892.797	−7363387.3	−7329465.5	−33894.815	−2145.8	−2146.4	−2.02	−29.607	−27.000
1.4	−7361241.3	−7327318.9	−33892.797	−7363262.0	−7329340.2	−33893.520	−2020.7	−2021.2	−0.72	−29.595	−28.324

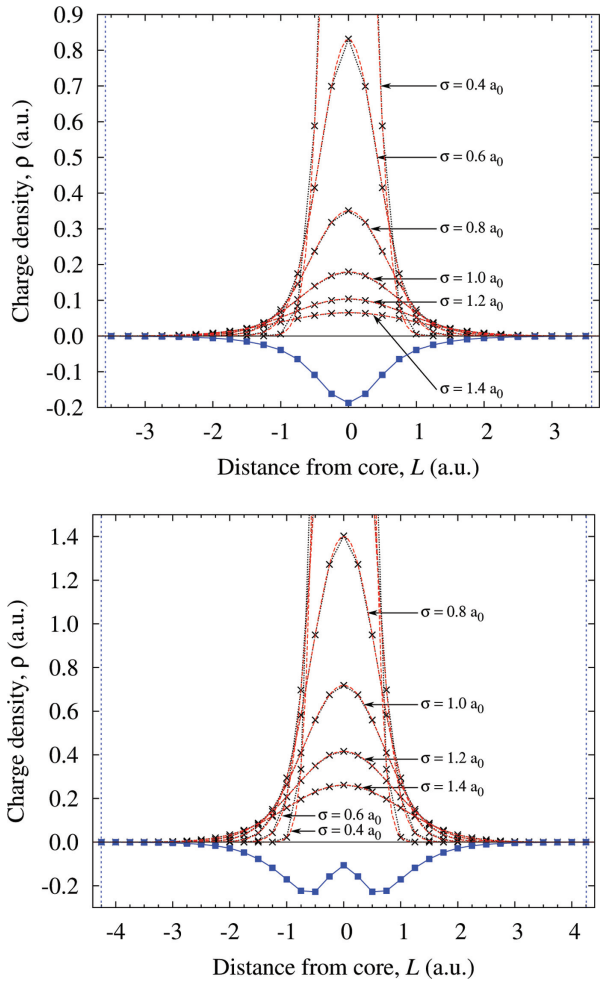

Figure 3. Cross-sections through charge densities for a hydrogen atom (top panel) and a carbon atom (bottom panel). Blue squares — electronic density. Black crosses—smeared ions with varying smearing widths. The spacing of the points in the plot coincides with the grid spacing used in the simulations. The dashed red curves indicate the exact shape of the Gaussians, whereas the dotted black lines denote a linear interpolation between the points. The vertical dashed lines correspond to the location of the cavity surface. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

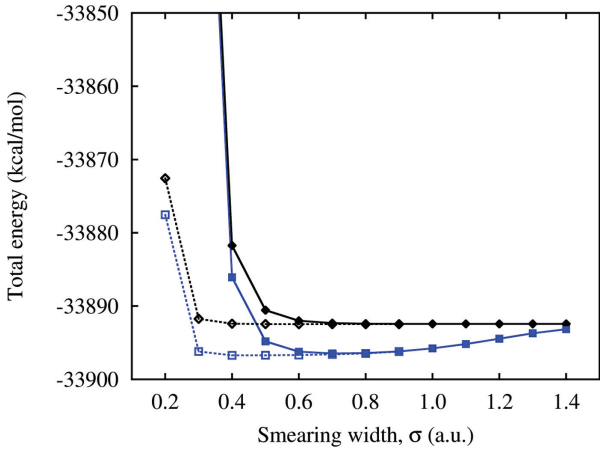

Figure 4. Total energy of the phenol ligand (black diamonds—in vacuum, blue squares—in solution) as a function of the smearing width used to smear the ions. Filled symbols and solid lines—for a grid with a spacing of 0.25 a_0. Empty symbols and dashed lines—for a grid with a spacing of 0.125 a_0. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

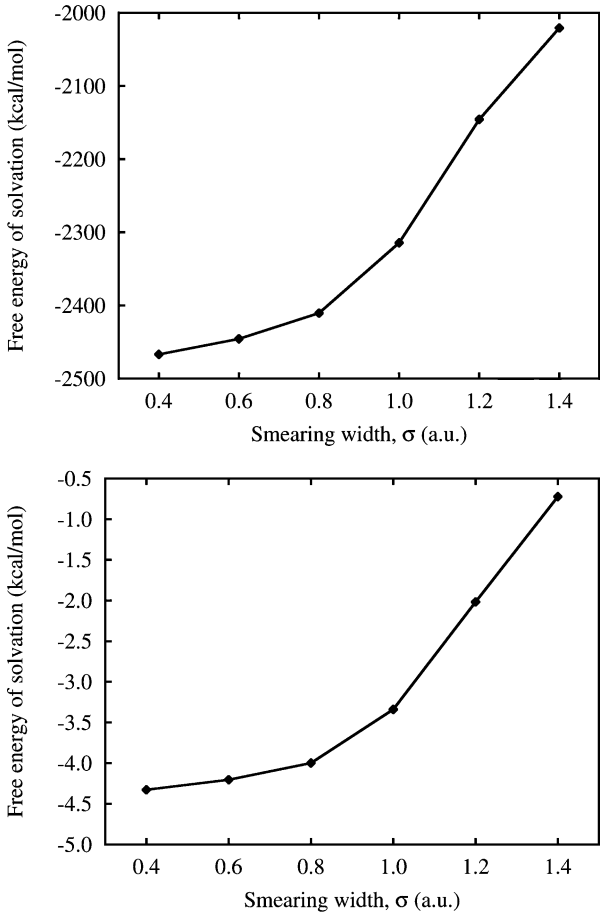

Figure 5. Free energy of solvation of the complex (top panel) and the phenol ligand (bottom panel) as a function of the smearing width used to smear the ions.

during the parameterization of the model used in this work, $\sigma = 0.8\,a_0$ was used throughout. Naturally, using the lowest smearing widths possible constitutes a more elegant solution to the problem but the associated necessity to make the grid finer quickly makes this impractical due to the inverse-cubic dependence of the CPU and memory requirements on the grid spacing.

We conclude this section by investigating the dependence of the ligand binding energy on the smearing width. As expected, the binding energy in vacuum is almost insensitive to the value of this parameter, as illustrated in Figure 7. The same figure shows the binding energy in solution, which, similarly to the free energy of solvation, rapidly deteriorates in accuracy when unphysically broad smearing widths are used. However, for "sensible," moderately broad smearing widths, such as $\sigma = 0.8\,a_0$, the effect on the accuracy is minimal. In this case, for example, the difference in the binding energy between calculations using

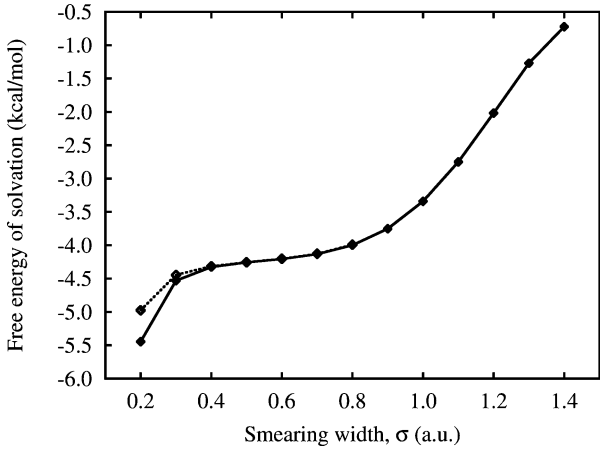

Figure 6. Free energy of solvation of the phenol ligand as a function of the smearing width used to smear the ions. Filled symbols and solid lines—for a grid with a spacing of 0.25 a_0. Empty symbols and dashed lines – for a grid with a spacing of 0.125 a_0.

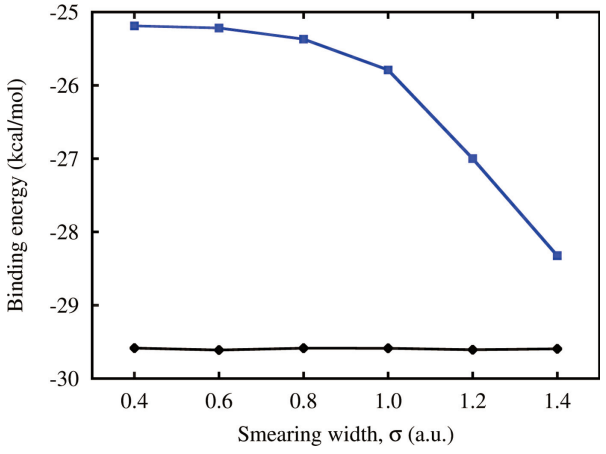

Figure 7. Binding energy of phenol in vacuum (black diamonds), and in solution (blue squares) as a function of the smearing width σ. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$\sigma = 0.8\,a_0$ and those using $\sigma = 0.4\,a_0$ was less than 0.2 kcal/mol or about 0.7%.

Approximations in the boundary conditions

Here, we investigate the effect of the approximations (4) and (5) on the accuracy of the obtained absolute energies, free energies of solvation and ligand binding energies. In each case, we shall use the value obtained for $p = 1$ (i.e., without any coarse-graining of the charge) as reference. In addition, we shall investigate the magnitude of the error incurred by using zero-Dirichlet boundary conditions instead of the open boundary conditions. As has been already mentioned in "Details of the DFT Calculations" section, only single calculations of the Coulombic energy (A6) have been performed here for the sake of reducing the computational effort. For this reason, further discussion will concern the polar component of the free energy of solvation rather than the free energy of solvation itself and the Coulombic component of the binding energy (i.e., the difference in Eq. (A6) in solvent and in vacuum) rather than the binding energy itself.

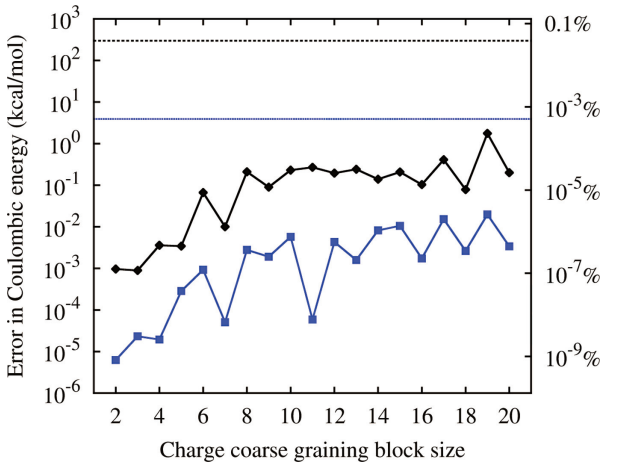

Figure 8. Error in the Coulombic energy of the complex (black diamonds—in vacuum, blue squares—in solution) as a function of the block size used for the calculation of boundary conditions. The horizontal lines correspond to the error incurred when zero boundary conditions are used instead (dashed black—in vacuum, dotted blue—in solution). The percentages reported on the right axis correspond to the Coulombic energy in vacuum. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 8 shows the error in the Coulombic energy of the complex incurred by the use of the approximation (4) (in vacuum) and (5) (in solution). The magnitude of this error is seen to be extremely small for both the calculations in vacuum and in solution, even if the coarse-graining proceeds over large blocks. The fact that the error incurred in solution is smaller by about two orders of magnitude is a consequence of the screening of the solute by the dielectric. We note that while using zero-boundary condition is a reasonable approximation in solution (leading to an error of less than 4 kcal/mol or 0.0005% in the energy), the same approximation is rather inaccurate in vacuum (leading to an error of about 300 kcal/mol, or 0.04% in the energy). This demonstrates the necessity of using physically sound boundary conditions for charged molecules for obtaining accurate absolute energies.

Figure 9 shows the error in the electrostatic (polar) term of the free energy of solvation incurred by the approximations in question. As the complex is a large, charged molecule, this term itself is rather large in magnitude—about −2800 kcal/mol, which means that extraordinary care needs to be taken to calculate it to chemical accuracy. In all cases, the error incurred by the use of charge coarse-graining was below 2 kcal/mol (less than 0.1%). We typically use $p = 5$, which is shown to incur only a modest error of 0.003 kcal/mol (or about 0.0001%), while reducing the computational effort of calculating the boundary conditions by a factor of $p^3 = 125$. When zero-boundary conditions were used, the error was unacceptably large (about 10%).

In Figure 10, we show the error in the Coulombic component of the binding energy of phenol incurred by the approximations in question. Even for the largest blocks, this error is negligible, particularly in solution. We point out that owing to judicious cancellation of errors, even if zero boundary conditions are used, accurate binding energies in solution can be obtained, even though the solvation energies suffer from about 10% error

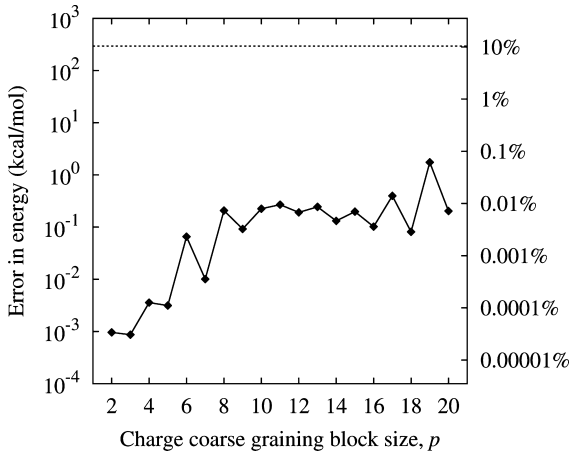

Figure 9. Error in the polar term of the free energy of solvation of the complex as a function of the block size used for the calculation of boundary conditions. The dashed line corresponds to the error incurred when zero boundary conditions are used instead.

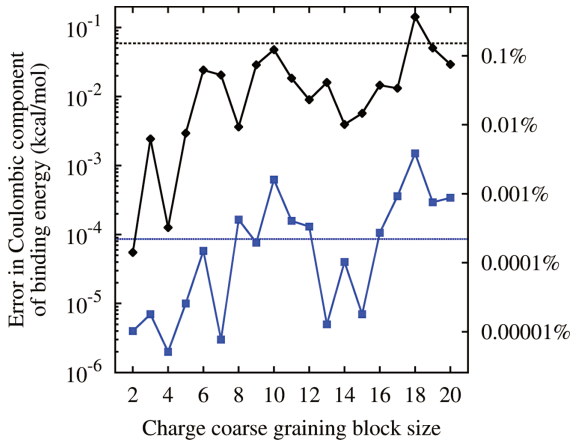

Figure 10. Error in the Coulombic component of the binding energy of phenol (black diamonds—in vacuum, blue squares—in solution) as a function of the block size used for the calculation of boundary conditions. The horizontal lines correspond to the error incurred when zero boundary conditions are used instead (dashed black—in vacuum, dotted blue—in solution). The percentages reported on the right axis correspond to the binding energy in vacuum. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

under this approximation. We stress that this should probably not be expected to hold in general—here, the ligand was small and neutral and the potential on the faces of the cell due to the complex and host was thus very similar. In consequence, the unphysical compensating potential introduced by zero boundary conditions was very similar in both cases, leading to good cancellation of errors. If the ligand was charged and/or larger, the degree of such cancellation would be much smaller. If open boundary conditions are used, this is no longer a concern.

Defect correction order

Here, we investigate the effect of using defect correction (cf. Appendix B) when solving the HPE (2) (in vacuum) or the NPE (1) (in solvent) and of the order of the finite differences used during defect correction on the accuracy of the obtained absolute energies, free energies of solvation, and ligand binding energies.

We begin by showing how *not* using defect correction and instead relying on a second-order solution of Eq. (2) introduces a non-negligible error into the Hartree energy of the phenol ligand in vacuum. We measure this error by comparison with two reference calculations performed with open boundary conditions. The first of these uses the Martyna–Tuckerman approach,[44] whereas the second one uses the cutoff-Coulomb technique[45] to calculate the Hartree energy with open boundary conditions. For a more detailed description of these approaches, the reader is referred to Ref. [15]. Our choice of the ligand for this test, rather than the larger lysozyme molecule, was dictated by the fact that significantly larger "padded" grids need to be used when using the aforementioned approaches, which increases their memory footprint. Although the Hartree energies computed with both reference approaches agreed to machine precision, the Hartree energy obtained by solving Eq. (2) (for $\rho_{\text{tot}} \equiv \rho$) with second-order finite differences and no defect correction differed by as much as 83 kcal/mol (about 0.07%), cf. Table 4. Using defect correction, we can drastically improve the agreement in the Hartree energy with our reference value. The magnitude of the discrepancy decreases monotonically with an increase in the order of the finite differences used, as shown in Figure 11. Smeared ions were not used in this case, as the quantity of interest, here, was the Hartree energy, rather than the Coulombic energy (A6), furthermore, we did not want the numerical inaccuracies of the smeared ions themselves to complicate the picture. The convergence criteria for the multigrid solver were identical in all cases, we assumed convergence when $|\phi^{(i+1)} - \phi^{(i)}| < 1 \times 10^{-5}$ (a.u.).

Table 4. Hartree energy (kcal/mol) of phenol in vacuum as a function of the finite-difference (FD) order used for the defect correction. An FD order of 2 corresponds to no defect correction.

FD order	Hartree energy in vacuum
2	113640.335
4	113558.656
6	113557.583
8	113557.502
10	113557.485
12	113557.480
MT	113557.47446894
CC	113557.47446894

MT and CC denote reference calculations performed with the Martyna–Tuckerman[44] and cutoff-Coulomb approaches,[45] respectively.

As neither of the methods used as a reference above could be applied in solution, and their use for large systems was cumbersome, in the remainder of this section, we shall use the results obtained with the highest finite difference order (i.e., 12), as a reference against which results obtained with lower orders or without defect correction will be compared. We will focus on the Coulombic energy (A6) as the relevant quantity in solution. Note, on the example of the phenol ligand, that this energy is much smaller in magnitude than the Hartree energy, as it

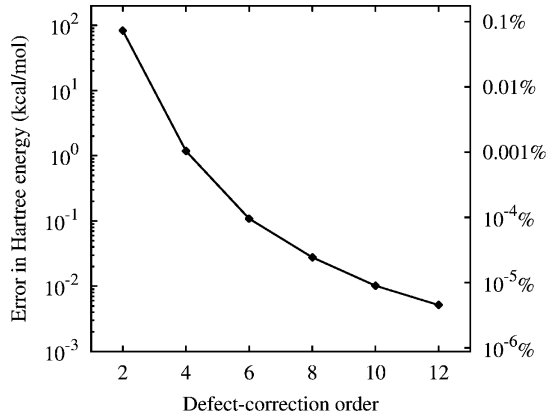

Figure 11. Convergence of the Hartree energy of phenol in vacuum to the exact result. Open boundary conditions were used. The exact result was obtained using the cutoff Coulomb technique and with the Martyna–Tuckerman approach, which agreed to machine precision.

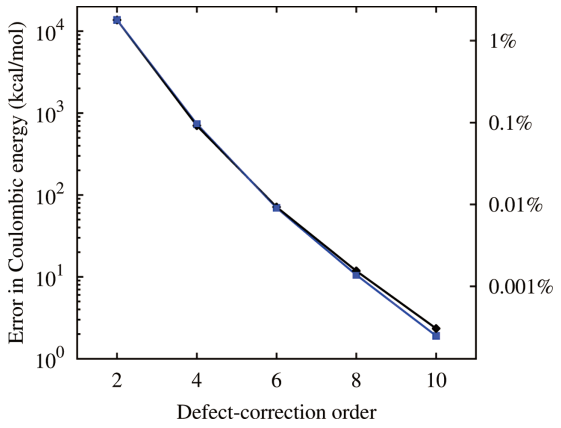

Figure 12. Error in the Coulombic energy of the complex (black diamonds—in vacuum, blue squares – in solution) as a function of the order of the finite differences used for the defect correction of the solution of the NPE. The order of 2 corresponds to no defect correction. The percentages reported on the right axis correspond to the energy in vacuum. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

corresponds to the interaction of a neutral charge distribution with itself, whereas the latter—to the interaction with itself of a distribution with a total charge of 36 electrons.

The obtained Coulombic energies, in vacuum and in solution, for the complex, host, and phenol ligand are shown in Table 5, whereas Table 6 lists the corresponding errors. For the case of the complex, this error is plotted in Figure 12. The behavior of the error is very similar in vacuum and in solution and across the studied systems. When defect correction is not used, errors as large as 1.7% in the Coulombic energy are incurred. This corresponds to more than 13,000 kcal/mol for the complex and host and about 73 kcal/mol for the ligand. When defect correction is used, the magnitude of the error quickly diminishes monotonically with increasing order of the finite differences used in the defect correction. This demonstrates the importance of using defect correction (or a high-order multigrid solver), if accurate absolute energies are desired.

We now shift our attention to the effect of defect correction on the free energies of solvation. Again, we shall only be concerned with the polar term, the values for which are shown in Table 7 for

Table 5. Coulombic energy (in kcal/mol) in vacuum and in solution of the complex, host, and phenol ligand, as a function of the order of the finite differences used for the defect correction of the solution of the Poisson equation.

	Coulombic energy in vacuum			Coulombic energy in solvent		
FD order	Complex	Host	Ligand	Complex	Host	Ligand
12	765280.2	761079.1	4240.088	762448.9	758252.6	4216.994
10	765282.5	761081.4	4240.098	762450.8	758254.5	4217.000
8	765292.0	761090.9	4240.141	762459.5	758263.1	4217.033
6	765351.4	761150.0	4240.432	762518.5	758321.8	4217.305
4	765983.6	761778.9	4243.688	763182.0	758982.2	4220.876
2	778986.6	774712.7	4312.981	776155.0	771885.9	4289.865

An order of 2 corresponds to no defect correction.

Table 6. Error in the Coulombic energy (in kcal/mol and as a percentage) in vacuum and in solution of the complex, host, and phenol ligand, with respect to the defect correction order of 12, as a function of the order of the finite differences used for the defect correction of the solution of the Poisson equation.

	Error in the Coulombic energy in vacuum						Error in the Coulombic energy in solvent					
FD order	complex		host		ligand		complex		host		ligand	
12	0.0	0.0000%	0.0	0.0000%	0.000	0.0000%	0.0	0.0000%	0.0	0.0000%	0.000	0.0000%
10	2.3	0.0003%	2.3	0.0003%	0.010	0.0002%	1.9	0.0002%	1.9	0.0002%	0.006	0.0001%
8	11.8	0.0015%	11.8	0.0015%	0.053	0.0013%	10.6	0.0014%	10.5	0.0014%	0.039	0.0009%
6	71.2	0.0093%	70.9	0.0093%	0.344	0.0081%	69.6	0.0091%	69.2	0.0091%	0.311	0.0074%
4	703.4	0.0919%	699.8	0.0920%	3.600	0.0849%	733.1	0.0962%	729.6	0.0962%	3.882	0.0921%
2	13706.4	1.7910%	13633.6	1.7913%	72.893	1.7191%	13706.1	1.7976%	13633.3	1.7980%	72.871	1.7280%

An order of 2 corresponds to no defect correction.

Table 7. Polar component of the free energy of solvation and the Coulombic component of the binding energy of the ligand in vacuum and in solution, as a function of the order of the finite differences used for the defect correction of the solution of the Poisson equation.

	Polar free energy of solvation			Coulombic component to binding energy	
FD order	Complex	Host	Ligand	In vacuum	In solution
12	−2831.3	−2826.5	−23.094	−39.001	−20.683
10	−2831.7	−2827.0	−23.098	−39.001	−20.678
8	−2832.5	−2827.8	−23.108	−39.001	−20.666
6	−2832.9	−2828.2	−23.127	−39.001	−20.640
4	−2801.6	−2796.7	−22.812	−39.001	−21.072
2	−2831.6	−2826.8	−23.116	−39.095	−20.806

The order of 2 corresponds to no defect correction. Energies in kcal/mol.

the complex, host, and the phenol ligand. The corresponding error and its dependence on the order of the finite differences is shown in Table 8 and in Figure 13. The magnitude of the error again decreases monotonically as the order of the finite differences used for defect correction is increased, however, interestingly, not using defect correction at all incurs only a negligible error in the free energy of solvation, even though the corresponding absolute energies are severely affected.

This somewhat counterintuitive observation is best explained by referring to Appendix B. It is a consequence of the approximate nature of Eq. (B6) used during defect correction: even though the residual $r_d^{(i)}$ is computed with high-order finite differences, the corresponding approximation of the algebraic error $e_{2,d}^{(i)}$ is still obtained with a second-order solver. When solving Eq. (B6), the solver deals with a right-hand side that is a discrete representation of $r_d^{(i)}$ on the grid, where normally (when solving Eq. (1) or (2)) the right-hand side is $-4\pi\rho_{\text{tot}}$. Although the magnitude of $r_d^{(i)}$ is significantly smaller than that of $-4\pi\rho_{\text{tot}}$ (as the former is merely the inaccuracy in the latter), the reverse is true for the gradients of these quantities (which we verified for the system in question). As

$$\begin{aligned} r^{(i)} &= f - \hat{A}\phi^{(i)} \\ &= f - \nabla\cdot(\varepsilon\nabla\phi^{(i)}) \\ &= f - (\nabla\varepsilon)\cdot(\nabla\phi^{(i)}) - \varepsilon\,\nabla^2\phi^{(i)}, \end{aligned} \tag{6}$$

the approximation inherent in Eq. (B6) becomes less accurate in the solvent calculation (where ε is nonhomogeneous) compared with the calculation in vacuum. Thus, even though the defect-corrected ϕ are more accurate than the uncorrected ones (both in vacuum and in solution), the fact that the quality of the approximation involved in defect correction differs between the calculations in vacuum and in solution translates into an additional error in the obtained free energy of solvation (their difference). In the absence of defect correction, this approximation is not being made at all, which leads to fortuitous cancellation of errors. As a further confirmation of this numerical behavior, we note that that the error in the free energy of solvation depends

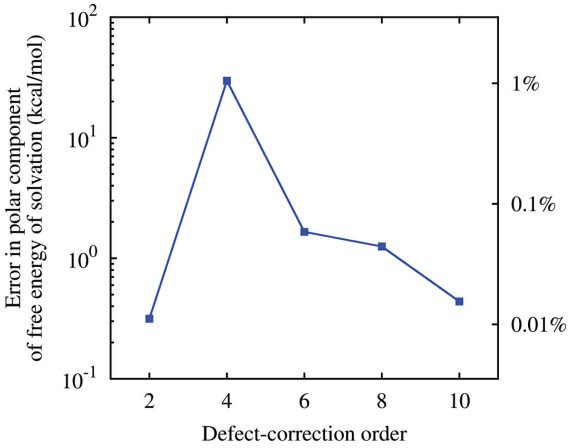

Figure 13. Error in the polar component of the free energy of solvation of the complex as a function of the order of the finite differences used for the defect correction of the solution of the NPE. An order of 2 corresponds to no defect correction. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 8. Error (in kcal/mol and as a percentage) in the polar term of the free energy of solvation and in the Coulombic component of the binding energy of the ligand in vacuum and in solution, with respect to the defect correction order of 12, as a function of the order of the finite differences used for the defect correction of the solution of the Poisson equation.

	Error in the polar free energy of solvation						Error in the Coulombic component of binding energy			
FD order	Complex		Host		Ligand		In vacuum		In solution	
12	0.0	0.000%	0.0	0.000%	0.000	0.000%	0.000	0.000%	0.000	0.000%
10	−0.4	0.015%	−0.4	0.016%	−0.004	0.019%	0.000	0.000%	0.005	−0.025%
8	−1.3	0.044%	−1.3	0.044%	−0.014	0.062%	0.000	0.000%	0.018	−0.085%
6	−1.7	0.058%	−1.7	0.059%	−0.033	0.143%	0.001	−0.001%	0.043	−0.210%
4	29.7	−1.049%	29.8	−1.055%	0.282	−1.222%	0.000	0.000%	−0.388	1.878%
2	−0.3	0.011%	−0.3	0.009%	−0.022	0.097%	−0.094	0.240%	−0.123	0.593%

An order of 2 corresponds to no defect correction.

on the smoothness of the dielectric function (which is controlled by the parameter β as presented in Ref. [5]). If we decrease β (leading to a decrease in the values of $\nabla\varepsilon$), we observe that the errors in the free energy of solvation obtained with low-order defect correction quickly decrease. In the case of the phenol ligand, for $\beta < 0.6$ even fourth-order defect correction leads to more accurate results than the uncorrected second-order calculation.

We conclude this section with a discussion of the effect of defect correction on the binding energy of phenol. The values of this binding energy, in vacuum and in solvent, for different orders of the finite differences used for defect correction are shown in Table 7 and the corresponding errors—in Table 8. These energies are also plotted in Figure 14, which demonstrates that in vacuum we deal with a convenient cancellation of errors, and the binding energy is not sensitive to more than 0.001 kcal/mol to the details of the defect correction. Even without defect correction, the error is smaller than 0.1 kcal/mol. The binding energy in solvent is somewhat more sensitive to this parameter, however, it still converges well as the order of the finite differences is increased. Nevertheless, defect correction with the lowest order (i.e., 4), introduces a larger error than neglecting defect correction altogether, for reasons similar to those discussed in the case of the free energy of solvation. In all cases, the obtained ligand binding energy differs less than 0.5 kcal/mol from the "exact" 12th order defect-corrected value.

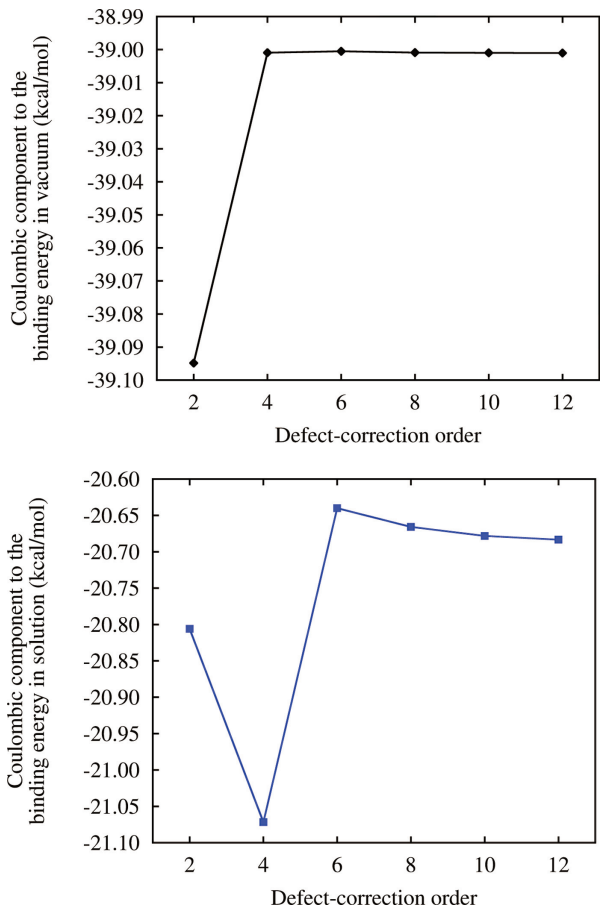

Figure 14. Coulombic component of the binding energy of phenol in vacuum (top panel) and in solvent (bottom panel) as a function of the order of the finite differences used for the defect correction of the solution of the Poisson equation. An order of 2 corresponds to no defect correction. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Conclusions

We have investigated the binding of two ligands (phenol, which is a polar ligand and toluene which is nonpolar) to the T4 lysozyme L99A/M102Q protein using large-scale DFT calculations, taking the solvent environment into account within the framework of a self-consistent implicit solvent model[7] with direct solution of the NPE, as implemented in the ONETEP program. We have investigated the behavior of the main numerical parameters that need to be carefully considered, when performing calculations of solvation energies and of binding energies in solvent.

A numerical parameter specific to the ONETEP linear-scaling DFT program is the localization radius of the NGWFs. We have shown that a localization radius of 8 a_0 for the NGWFs is sufficient to accurately obtain free energies of solvation and the binding energies for both ligands. We have demonstrated that the binding energies can only be accurately obtained if dispersion is explicitly taken into account, for example, with the DFT+D approach[39] in the case of the GGA exchange-correlation functionals, as we use here.

To accurately describe the polarization of the implicit solvent due to the potential of the ionic cores, we used the smeared-ion formalism. We have shown how the width of the smearing needs to be carefully chosen, otherwise the energies in solution, and consequently, free energies of solvation and binding energies will be adversely affected. We have ensured that our calculations are converged with respect to this parameter. We propose to use 0.8 a_0 for the smearing width, unless grids substantially finer than 0.25 a_0 are used.

Throughout this work, we used open boundary conditions. We used coarse-graining when generating the open boundary conditions for the solution of the Poisson equation, both in vacuum [Eq. (2)] and in solvent [Eq. (1)]. We have shown that the effect on accuracy associated with this approximation is negligible, even when very crude representations of the charge density are used. We demonstrated that using zero-boundary conditions, when solving the Poisson equation incurs unacceptably large errors in vacuum, but is a reasonable approximation in solvent, owing to the dielectric screening.

We have investigated how the order to which the Poisson equation is solved affects the obtained total energies, free energies of solvation, and binding energies. It has been pointed out[19,46] that second-order finite differences are often inadequate in the context DFT calculations. In this work, we have used the defect correction technique,[18,19] which serves to accurately approximate a high-order solution using a second-order solver and high-order finite difference operators for the gradient and Laplacian. We have shown how a second-order solution leads to highly inaccurate total energies, yet a high degree of cancellation of errors allows one to obtain rather accurate free

energies of solvation without using defect correction. We have demonstrated that defect correction offers excellent improvement in the accuracy of the obtained total energies, but because of the approximations involved, in the case of the free energies of solvation an increase in accuracy is only obtained when high-order (10th or so) finite differences are used, depending on the exact nature of the system under study. For the binding energies, we have shown that high-order defect correction offers better accuracy, however, even without defect correction the error in the binding energy was smaller than 0.2 kcal/mol.

We hope that our case study provides a detailed understanding of the strengths and weaknesses of such minimal parameter solvent models and will hence contribute to their optimal usage.

Appendix A: Electrostatics in the presence of smeared ions

We will be concerned with the electrostatic energy of an isolated (nonperiodic) system of N ionic cores and N_e electrons. We will follow a convention where electronic charge densities are positive whereas the cores are negatively charged. Let $\rho_I(\mathbf{r})$ be a Gaussian charge density centerd on core I that integrates to the core's charge, $-Z_I$, that is,

$$\rho_I(\mathbf{r}) = -\frac{Z_I}{\sigma_I^3}\pi^{-\frac{3}{2}} \exp\left(-\frac{|\mathbf{r}-\mathbf{R}_I|^2}{\sigma_I^2}\right). \tag{A1}$$

We denote the potential due to the density (A1) with $v_I(\mathbf{r})$:

$$\begin{aligned} v_I(\mathbf{r}) &= \int \frac{\rho_I(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}\mathbf{r}' \\ &= -\frac{Z_i}{|\mathbf{r}-\mathbf{R}_I|}\,\mathrm{erf}\left(\frac{|\mathbf{r}-\mathbf{R}_I|}{\sigma_I}\right). \end{aligned} \tag{A2}$$

The total density $\rho_{\mathrm{tot}}(\mathbf{r})$ of the molecule is the sum of the electronic density $\rho(\mathbf{r})$ and Gaussian core densities

$$\rho_{\mathrm{tot}}(\mathbf{r}) = \rho(\mathbf{r}) + \sum_I^N \rho_I(\mathbf{r}). \tag{A3}$$

It can be shown[47] that solution of the NPE (Eq. (1)) for the above density results in an electrostatic potential of the following form

$$\phi_{\mathrm{NPE}}(\mathbf{r}) = \phi_{\mathrm{HPE}}(\mathbf{r}) + \phi_{\mathrm{pol}}(\mathbf{r}), \tag{A4}$$

where the polarization potential $\phi_{\mathrm{pol}}(\mathbf{r})$ contains all the effects due to the nonhomogeneous dielectric permittivity ε in Eq. 1 and $\phi_{\mathrm{HPE}}(\mathbf{r})$ is the solution of the HPE [Eq. (2)].

The electrostatic energy of $\rho_{\mathrm{tot}}(\mathbf{r})$ interacting with itself in the presence of the dielectric is then

$$E_{\mathrm{tot}} = \frac{1}{2}\int \rho_{\mathrm{tot}}(\mathbf{r})\,\phi_{\mathrm{NPE}}(\mathbf{r})\,d\mathbf{r} \tag{A5}$$

$$= \frac{1}{2}\int \rho_{\mathrm{tot}}(\mathbf{r})\,\phi_{\mathrm{HPE}}(\mathbf{r})\,d\mathbf{r} + \frac{1}{2}\int \rho_{\mathrm{tot}}(\mathbf{r})\,\phi_{\mathrm{pol}}(\mathbf{r})\,d\mathbf{r}$$

$$= E_{\mathrm{HPE}} + E_{\mathrm{pol}}, \tag{A6}$$

where we have conceptually partitioned the energy into homogeneous E_{HPE} and polarization E_{pol} terms. In practice, however, we compute directly the sum of the two above terms E_{tot}, as in the case of solvent, we only solve the NPE.

The actual electrostatic energy that we compute in the DFT calculation of a solute molecule in implicit solvent has the following form:

$$E_{\mathrm{ES,DFT}} = \frac{1}{2}\int \rho_{\mathrm{tot}}(\mathbf{r})\,\phi_{\mathrm{NPE}}(\mathbf{r})\,d\mathbf{r} \tag{A7}$$

$$+ \sum_I^N \int \rho(\mathbf{r})[v_{\mathrm{loc},I}(\mathbf{r}) - v_I(\mathbf{r})]\,\mathrm{d}\mathbf{r} \tag{A8}$$

$$- \frac{1}{2}\sum_{I,J}^N \iint \frac{\rho_I(\mathbf{r})\rho_J(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}' \tag{A9}$$

$$+ \frac{1}{2}\sum_{\substack{I,J \\ I\neq J}}^N \frac{Z_I Z_J}{|\mathbf{R}_I-\mathbf{R}_J|}, \tag{A10}$$

where the term A8 corrects the E_{HPE} component of E_{tot} for the fact that in the molecule the interaction of the electronic density is with local pseudopotentials rather than the smeared Gaussian nuclei, and the Gaussian–Gaussian interaction is removed from E_{HPE} in Eq. (A9) and replaced with the interaction between the pseudopotential cores in Eq. (A10) which under the usual (valid) assumption of nonoverlap of the ionic cores is the Coulombic interaction between point ionic charges.

Using standard formulas,[48] the Gaussian–Gaussian term of Eq. (A9) is computed as

$$-\frac{1}{\sqrt{2\pi}}\sum_I^N \frac{Z_I^2}{\sigma_I} - \frac{1}{2}\sum_{\substack{I,J \\ I\neq J}}^N \frac{Z_I Z_J}{|\mathbf{R}_I-\mathbf{R}_J|}\,\mathrm{erf}\left(\frac{|\mathbf{R}_I-\mathbf{R}_J|}{\sqrt{\sigma_I^2+\sigma_J^2}}\right). \tag{A11}$$

The only correction term we have neglected from $E_{\mathrm{ES,DFT}}$ from the above equation is the following

$$\frac{1}{2}\sum_I^N \int [-Z_I - \rho_I(\mathbf{r})]\phi_{\mathrm{pol}}(\mathbf{r})d\mathbf{r},$$

which would remove the interaction of the smeared Gaussian cores with the polarization potential and replace it with the interaction of the point ionic charges with the polarization potential. Even though we do not apply this correction, as we do not have access to $\phi_{\mathrm{pol}}(\mathbf{r})$, we expect its effect to be negligible. It can be shown[47] that the polarization potential can be expressed as the potential due to a polarization density $\rho_{\mathrm{pol}}(\mathbf{r})$, which is localized to a narrow region along the vacuum–solute interface (the effective surface of the solute cavity). The above term can thus be rewritten as

$$\frac{1}{2}\sum_I^N \int \left[\frac{-Z_I}{|\mathbf{r}-\mathbf{R}_I|} - v_I(\mathbf{r})\right]\rho_{\mathrm{pol}}(\mathbf{r})d\mathbf{r}.$$

As the potential (A2) due to a Gaussian charge distribution quickly tends to that of a point charge, the difference between the two

(and, consequently, the correction term above) is expected to be negligible at that interface.

Obviously, in the case of calculations in vacuum all the terms with the subscript "pol" are zero by definition.

Appendix B: High-order defect correction

Here, we will briefly describe how a high-order defect correction method can be applied to reduce the error resulting from the second-order representation of the differential operators when solving the NPE (1) with a second-order multigrid solver. For a more detailed description of this approach, we refer the reader to Refs. [18, 19].

Our goal is to numerically solve Eq. (1), which we rewrite as

$$\hat{A}[\varepsilon]\phi = f, \tag{B1}$$

where $\hat{A}[\varepsilon] = \nabla \cdot \varepsilon\nabla$, $f = -4\pi\rho_{\text{tot}}$, and we have omitted the dependence on $\mathbf{r}$ for brevity.

Let $\hat{A}_d^h[\varepsilon]$ be a discrete representation of order d of $\hat{A}[\varepsilon]$ on a grid with a spacing of h. This representation is thus accurate to $O(h^d)$. For simplicity, we shall from now on omit the dependence of $\hat{A}$ and its discrete representations on ε from the notation.

The procedure which we describe here is an iterative improvement which is applied once the solution of Eq. (B1) with a second-order solver has been obtained. It is clear that the solution of Eq. (B1) with a second-order solver does not yield the sought ϕ, but rather $\phi^{(1)}$, which satisfies

$$\hat{A}_2^h\phi^{(1)} = f \tag{B2}$$

and is the starting guess for our iterative improvement. The difference between ϕ and $\phi^{(i)}$, at any iteration i of the calculation, is termed the algebraic error, $e^{(i)}$:

$$e^{(i)} = \phi - \phi^{(i)}. \tag{B3}$$

By applying $\hat{A}$ to both sides of (B3), we obtain the so-called defect equation:

$$\begin{aligned}\hat{A}e^{(i)} &= \hat{A}\phi - \hat{A}\phi^{(i)} \\ &= f - \hat{A}\phi^{(i)} \\ &= r^{(i)},\end{aligned} \tag{B4}$$

where $r^{(i)} = f - \hat{A}\phi^{(i)}$ is termed the residual (or defect) at iteration i.

Although, as we lack the exact $\hat{A}$, we can never exactly compute the residual, we can use a high-order approximation to $\hat{A}$ to compute the approximation to the residual $r_d^{(i)}$, of order d, as

$$r_d^{(i)} = f - \hat{A}_d^h\phi^{(i)}. \tag{B5}$$

We can then use the second-order analog of the defect equation (B4) to obtain a second-order approximation of the algebraic error, $e_{2,d}^{(i)}$, corresponding to this approximation to the residual, that is, we solve

$$\hat{A}_2^h e_{2,d}^{(i)} = r_d^{(i)} \tag{B6}$$

with the second-order solver. From Eq. (B3), it follows that

$$\begin{aligned}\phi &= \phi^{(i)} + e^{(i)} \\ &\approx \phi^{(i)} + e_{2,d}^{(i)}.\end{aligned} \tag{B7}$$

We can thus obtain a better approximation of the sought quantity ϕ as

$$\phi^{(i+1)} = \phi^{(i)} + e_{2,d}^{(i)} \tag{B8}$$

using only a second-order solver and a high-order representation of the operator $\nabla \cdot \varepsilon\nabla$.

This procedure is iterated until an appropriate convergence criterion is satisfied, for example, $|\phi^{(i+1)} - \phi^{(i)}|$ is below a prescribed tolerance.

Acknowledgments

J.D. acknowledges the support of the Engineering and Physical Sciences Research Council (EPSRC grant No. EP/G-55882/1) and of the Polish Ministry of Science and Information Technology (grant N N519 577838). S.J.F. would like to thank the BBSRC and Boehringer Ingelheim 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 galera supercomputer at the TASK Computer Centre (Gdansk, Poland).

Keywords: DFT • implicit solvent • ONETEP

How to cite this article: J. Dziedzic, S. J. Fox, T. Fox, C. S. Tautermann, C.-K. Skylaris, *Int. J. Quantum Chem.* **2013**, *113*, 771–785. DOI: 10.1002/qua.24075

[1] J. Tomasi, M. Persico, *Chem. Rev.* **1994**, *94*, 2027.
[2] A. Klamt, G. Schüürmann, *J. Chem. Soc. Perkin Trans.* **1993**, *2*, 799.
[3] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Chem. Theor. Comput.* **2009**, *5*, 2447.
[4] B. Mennucci, E. Cancè, J. Tomasi, *J. Phys. Chem. B* **1997**, *101*, 10506.
[5] J.-L. Fattebert, F. Gygi, *J. Comp. Chem.* **2002**, *23*, 662.
[6] D. Scherlis, J. Fattebert, F. Gygi, M. Cococcioni, N. Marzari, *J. Chem. Phys.* **2006**, *124*, 074103.
[7] J. Dziedzic, H. H. Helal, C.-K. Skylaris, A. A. Mostofi, M. C. Payne, *Europhys. Lett.* **2011**, *95*, 43001.
[8] A. Marenich, C. Kelly, J. Thompson, G. Hawkins, C. Chambers, D. Giesen, P. Winget, C. Cramer, D. Truhlar, Minnesota solvation database, version 2009, University of Minnesota: Minneapolis, 2009.
[9] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper and V. S. Pande, *J. Med. Chem.* **2008**, *51*, 769.
[10] J. P. Guthrie, *J. Phys. Chem. B* **2009**, *113*, 4501.
[11] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, M. C. Payne, *J. Chem. Phys.* **2005**, *122*, 084119.
[12] P. Hohenberg, W. Kohn, *Phys. Rev.* **1964**, *136*, 864.

[13] W. Kohn, L. J. Sham, *Phys. Rev.* **1965**, *140*, 1133.
[14] N. Hadjisavvas, A. Theophilou, *Phys. Rev. A* **1984**, *30*, 2183.
[15] N. D. M. Hine, J. Dziedzic, P. D. Haynes, C.-K. Skylaris, *J. Chem. Phys.* **2011**, *135*, 204103.
[16] A. Brandt, *Math. Comput.* **1977**, *31*, 333.
[17] M. P. Merrick, K. A. Iyer, T. L. Beck, *J. Phys. Chem.* **1995**, *99*, 12478.
[18] S. Schaffer, *Math. Comput.* **1984**, *43*, 89.
[19] H. Helal, *Including Solvent Effects in First-Principles Simulations of Biological Systems*, Ph.D. thesis, University of Cambridge, 2010.
[20] E. Baldwin, W. A. Baase, X. Zhang jun, V. Feher, B. W. Matthews, *J. Mol. Biol.* **1998**, *277*, 467.
[21] S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill, B. K. Shoichet, *J. Mol. Biol.* **2009**, *394*, 747.
[22] W. A. Baase, X. J. Zhang, D. W. Heinz, M. Blaber, E. P. Baldwin, B. W. Matthews, A. E. Eriksson, *Science* **1992**, *255*, 178.
[23] A. E. Eriksson, W. A. Baase, B. W. Matthews, *J. Mol. Biol.* **1993**, *229*, 747.
[24] A. Morton, W. Baase, B. W. Matthews, *Biochemistry* **1995**, *34*, 8564.
[25] B. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, B. K. Shoichet, *J. Mol. Biol.* **2002**, *322*, 339.
[26] B. Q. Wei, L. H. Weaver, A. M. Ferrari, B. W. Matthews, B. K. Shoichet, *J. Mol. Biol.* **2004**, *337*, 1161.
[27] P. Labute, *Proteins* **2009**, *75*, 187.
[28] MOE2009.10, Chemical Computing Group: Montreal, **2009**.
[29] D. Case, T. Darden, T. C., III, C. Simmerling, J. Wang, R. Duke, R. Luo, M. Crowley, W. Zhang, K. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvai, K. Wong, F. Paesani, J. Vanicek, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. Mathews, M. Seetin, C. Sagui, V. Babin, R. C. Walker, P. Kollman, University of California, San Francisco, *AMBER 10*, **2008**.
[30] J.-P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327.
[31] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, *Proteins: Struct, Funct Gen* **2006**, *65*, 712.
[32] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, *J. Chem. Phys.* **1983**, *79*, 926.
[33] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comp. Chem.* **2004**, *25*, 1157.
[34] J. Srinivasan, T. E. Cheatham, III, P. Cieplak, P. A. Kollman, D. A. Case, *J. Am. Chem. Soc.* **1998**, *120*, 9401.
[35] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, T. E. Cheatham, *Acc. Chem. Res.* **2000**, *33*, 889.
[36] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez, M. C. Payne, *Phys. Rev. B* **2002**, *66*, 035119.
[37] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris, M. C. Payne, *J. Chem. Phys.* **2003**, *119*, 8842.
[38] P. D. Haynes, C.-K. Skylaris, A. A. Mostofi, M. C. Payne, *Chem. Phys. Lett.* **2006**, *422*, 345.
[39] Q. Hill and C.-K. Skylaris, *Proc. R. Soc. A* **2009**, *465*, 669.
[40] S. J. Fox, C. Pittock, T. Fox, C. Tautermann, N. Malcolm, C.-K. Skylairs, *J. Chem. Phys.* **2011**, *135*, 224017.
[41] C.-K. Skylaris, O. Diéguez, P. D. Haynes, M. C. Payne, *Phys. Rev. B* **2002**, *66*, 073103.
[42] G. Román-Pérez, J. Soler, *Phys. Rev. Lett.* **2009**, *103*, 096102.
[43] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.
[44] G. J. Martyna, M. E. Tuckerman, *J. Chem. Phys.* **1999**, *110*, 2810.
[45] M. R. Jarvis, I. D. White, R. W. Godby, M. C. Payne, *Phys. Rev. B* **1997**, *56*, 14972.
[46] J. R. Chelikowsky, N. Troullier, Y. Saad, *Phys. Rev. Lett.* **1994**, *72*, 1240.
[47] O. Andreussi, I. Dabo, N. Marzari, *J. Chem. Phys.* **2012**, *136*, 064102.
[48] Szabo, N. S. Ostlund, *Modern Quantum Chemistry*, McGraw Hill: New York, **1982**.

Received: 15 December 2011
Revised: 7 February 2012
Accepted: 10 February 2012
Published online on 28 March 2012