

This is the peer reviewed version of the following article:

Fox S., Dziedzic J., Fox T., Tautermann C., Skylaris C.: Density functional theory calculations on entire proteins for free energies of binding: Application to a model polar binding site, *PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS*, Vol. 82, Iss. 12 (2014), pp. 3335-3346,

which has been published in final form at <https://doi.org/10.1002/prot.24686>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley's version of record on Wiley Online Library and any embedding, framing or otherwise making available the article or pages thereof by third parties from platforms, services and websites other than Wiley Online Library must be prohibited.

## DFT calculations on entire proteins for free energies of binding: application to a model polar binding site

Running title: Free energies of binding from DFT calculations on entire proteins

Stephen Fox<sup>1</sup>, Jacek Dziedzic<sup>1,3</sup>, Thomas Fox<sup>2</sup>, Christofer S. Tautermann<sup>2</sup>, and Chris-Kriton Skylaris<sup>1\*</sup>

<sup>1</sup>School of Chemistry, University of Southampton, Southampton, Hampshire SO17 1BJ, UK

<sup>2</sup>Boehringer Ingelheim Pharma GmbH & Co KG, Department for Lead Identification and Optimization Support, 88397 Biberach, Germany. <sup>3</sup>Faculty of Applied Physics and Mathematics, Gdansk University of Technology, Gdansk, Poland

Corresponding author email address: C.Skylaris@soton.ac.uk

Key words: Free energies of binding, Protein-ligand interactions, DFT, Large-scale DFT, ONETEP, QM-PBSA, T4-lysozyme L99A/M102Q

## Abstract

In drug optimisation calculations, the Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) method can be used to compute free energies of binding of ligands to proteins. The method involves the evaluation of the energy of configurations in an implicit solvent model. One source of errors is the force field used, which can potentially lead to large errors due to the restrictions in accuracy imposed by its empirical nature. To assess the effect of the force field on the calculation of binding energies, in this paper we use large-scale Density Functional Theory (DFT) calculations as an alternative method to evaluate the energies of the configurations in a “QM-PBSA” approach. Our DFT calculations are performed with a near-complete basis set and a minimal parameter implicit solvent model, within the self-consistent calculation, using the ONETEP program on protein-ligand complexes containing more than 2600 atoms. We apply this approach to the T4-lysozyme double mutant L99A/M102Q protein, which is a well-studied model of a polar binding site, using a set of eight small aromatic ligands. We observe that there is very good correlation between the MM and QM binding energies in vacuum but less so in the solvent. The relative binding free energies from DFT are more accurate than the ones from the MM calculations, and give markedly better agreement with experiment for 6 out of the 8 ligands. Furthermore, in contrast to MM-PBSA, QM-PBSA is able to correctly predict a non-binder.

# 1 Introduction

Advances in computational chemistry and biochemistry are directed towards more accurate descriptions of protein-ligand binding energies, which are essential for the prediction of ligand binding affinities, a long-standing goal in the field of computational drug design [1, 2]. Many methods have been developed to tackle this problem, ranging from theoretically rigorous approaches such as free energy perturbation [3, 4] and thermodynamic integration (TI) [5], to cheap and fast scoring methods [6] commonly used for docking calculations.

A method of medium computational effort is the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) approach [7]. This approach combines molecular dynamics (MD) simulations and continuum solvation models to estimate protein-ligand binding free energies. A significant source of error in MM-PBSA can be the accuracy of the interaction energies computed for each snapshot, as this accuracy depends on the selected force field. Force fields have limitations due to their empirical parametrisation, which can be unreliable for novel compounds significantly different from those present in the fitting set. Another limitation of most force fields is their inability to explicitly describe electronic polarisation and charge transfer. To investigate the effect of these force field-based limitations, in this work we replace the MM energy with a first-principles quantum mechanical (QM) energy evaluated for the entire system, to perform QM-PBSA. For the QM calculations we use the linear-scaling DFT program ONETEP [8]. In a previous paper [9], we investigated the numerical parameters of the implicit solvation model within ONETEP [10], using as a test system the T4 lysozyme double mutant Leu99Ala/Met102Gln [11] (or L99A/M102Q). In this study we will also be using T4 lysozyme L99A/M102Q to calculate QM-PBSA-based relative free energies of binding for various ligands.

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 [12, 13]. Two well studied mutants of T4 lysozyme are Leu99Ala (L99A) [12, 14, 15, 16] and L99A/M102Q [11]. These mutations create a small buried apolar and polar cavity respectively, which are capable of encapsulating small aromatic ligands. These T4 lysozyme mutants have been used to compare and validate binding free energy methods [14, 17, 18, 19, 20, 21] and to develop docking procedures [11, 22, 23]. The relative simplicity of the T4 lysozyme mutants and their small size make them attractive for validating computational methods. Coupled with the abundance of literature, this makes T4 lysozyme L99A/M102Q a good choice for a benchmark of



QM-PBSA calculations.

In this study we are comparing the free energy of binding between the conventional MM-PBSA approach and our QM-PBSA approach on 8 ligands bound to the T4 lysozyme double mutant L99A/M102Q. These ligands were chosen as they comprise a variety of chemical and physical properties (polarity, inclusion of halides, size and non-binders). Section 2 describes the MM-PBSA method, some of its most common variants, ONETEP and our solvation model, QM-PBSA and our simulation protocols and parameters. In Section 3 we present and discuss our calculation results and, we finish with conclusions in section 4.

## 2 Methods

### 2.1 MM-PBSA and QM-PBSA

Ligand binding affinity is calculated as

$$\Delta G_{\text{bind}} = G_{\text{complex}} - G_{\text{receptor}} - G_{\text{ligand}}. \quad (1)$$

Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) [7] is a method for computationally predicting ligand binding affinity. The approach is based on the post-processing of a molecular dynamics trajectory, typically run in explicit solvent and counterions in a periodic box. The free energy of binding is estimated by extracting a representative structural ensemble of “snapshots” from the trajectory. Solvent molecules and counterions are removed, then MM is used to calculate the gas phase interaction energy and a continuum solvent model (PBSA) is used to calculate the solvation energies. The free energy of binding is then obtained as the average over the ensemble of structures:

$$\Delta G = \langle \Delta E^{\text{MM}} \rangle + \langle \Delta G_{\text{PBSA}} \rangle - T \langle \Delta S^{\text{MM}} \rangle, \quad (2)$$

where  $\langle \Delta E^{\text{MM}} \rangle$  is the interaction energy in vacuum,  $\langle \Delta G_{\text{PBSA}} \rangle$  is the solvation energy, and  $\langle \Delta S^{\text{MM}} \rangle$  is the solute entropy, which can be obtained using normal mode analysis. By using a continuum solvent model, the problem is simplified since we are implicitly integrating out all the solvent coordinates, which results in more rapid convergence with the number of snapshots.

The straightforward way to calculate the binding free energy is the three-trajectory approach, where separate simulations are carried out for the complex, receptor and ligand. However, it has been found that the one-trajectory approach, where only the complex simulation is run, and receptor and ligand configurations are extracted from the complex geometries, is more accurate due to error cancellation [24]. It is also 2-3 times faster, since the most computationally demanding part is the MD simulations. However, this approach assumes that there are no conformational changes to the ligand or receptor upon binding. Furthermore, when using the single trajectory approach,  $\langle E^{\text{MM}} \rangle$  is the difference in non-bonded terms only, since all bonded terms will cancel.

Calculation of the entropy in a consistent and accurate manner is challenging. This makes calculation of the absolute binding free energies difficult. An approximation that is often used is that entropy change is assumed to be comparable for similar ligands, and hence cancels when considering relative free energies of binding. Although this may seem a poor assumption, calculating the entropy of the ligand from structures taken from the complex trajectory may well be an equally poor simplification; since the ligand would be expected to have many more degrees of freedom when free in solution. The relative free energy is calculated as

$$\Delta\Delta G_{A\rightarrow B} = \Delta\langle\Delta E^{\text{MM}}\rangle_{A\rightarrow B} + \Delta\langle\Delta G_{\text{solv}}\rangle_{A\rightarrow B}. \quad (3)$$

A significant source of error in MM-PBSA can be the accuracy of the interaction energies computed for each snapshot, as this accuracy depends on the parametrisation and transferability of the selected force field. Apart from the obvious approach of using more advanced force fields, a related direction for improvement is to replace either part or all of the force field description by a quantum description of the system. This would be expected to be more accurate and transferable due to explicitly accounting for the electronic effects, which are the source of all the interactions.

Kaukonen *et al.* [25] presented a QM/MM-PBSA approach and compared it to MM-PBSA for the purpose of studying reactions in proteins. The aim was to study the stability of two states with a shared proton. The QM calculations were performed with DFT using the BP86 exchange-correlation functional with a 6-31G\* basis set and DZP for metal ions for one system, and the B3LYP functional for another system using the same basis sets. This method showed improved results for the proton transfer and was in good agreement with more rigorous approaches (QTCP [26]) with median absolute deviations (MADs) of 4-22 kJ/mol. However, the QM region was quite small, comprising 46 atoms while the MM part



comprised 12132 atoms.

Wang *et al.* [27] used the SIESTA DFT code [28] combined with the implicit solvation model in the UHBD software [29] in a QM/MM-PBSA approach. The only differences between the ligands in the pocket was a single chemical functional group, with all common atoms being in identical positions. This was done to improve the odds of cancellation of systematic errors when comparing binding free energies. Relaxed structures were generated in three different ways. The first was geometry optimised with SIESTA and the second and third in a QM/MM approach with the ONIOM method in GAUSSIAN03 [30]. However, no protein configurational sampling was performed. Wang *et al.* used a fixed geometry (single structure) approximation, where only minimised crystal structures are used.

Diaz *et al.* [31] replaced  $\Delta\langle\Delta E^{\text{MM}}\rangle$  with energies from linear-scaling semi-empirical QM calculations on an ensemble of structures from a classical MD simulation in a QM-PBSA type model. Prior to the single-point energy calculations, QM/MM geometry optimisations of a subsystem of the enzyme were performed, keeping the rest of the enzyme fixed. Single point energy calculations were performed with AM1 and PM3 using the divide and conquer (D&C) approach on the subsystem of the enzyme. The DivCon99 [32] program was used to perform the D&C calculations. Diaz *et al.* found that the resulting QM/MM geometry-optimised structures were similar to the MD representations generated from the force fields, and using semi-empirical QM D&C gave comparable relative binding free energies to MM-PBSA. However, this uses an empirical method and suffers from some of the same transferability problems as force fields.

Cole *et al.* [33] have recently extended the MM-PBSA approach to a full QM-PBSA approach, with sampling of protein motion, where the calculation of the interaction energies in vacuum by the force field is replaced by DFT calculations on the entire molecule for an ensemble of snapshots taken from an MD simulation. The energy of each snapshot is obtained as  $E^{\text{QM}} = E_{\text{DFT}} + E_{\text{disp}}$ , where  $E_{\text{disp}}$  is the dispersion correction [34] to the total DFT energy,  $E_{\text{DFT}}$ . In previous work [33, 35], the free energy of solvation in the QM calculation,  $G_{\text{solv}}^{\text{QM}}$  was obtained by scaling the classical solvation energy by the QM electrostatic energy, giving the free energy of binding as

$$\Delta G_{\text{tot}} = \langle\Delta E^{\text{QM}}\rangle + \langle\Delta G_{\text{PB}}\left(\frac{\Delta E_{\text{DFT}}}{\Delta E_{\text{EL}}}\right) + \Delta G_{\text{SA}}\rangle \quad (4)$$

$$= \langle\Delta E^{\text{QM}}\rangle + \langle\Delta G_{\text{solv}}^{\text{QM}}\rangle, \quad (5)$$

where  $\Delta E_{\text{EL}}$  is the electrostatic contribution to the binding energy from the MM calculation,  $\Delta G_{\text{PB}}$  is the polar term from the solvation energy and  $\Delta G_{\text{SA}}$  is the non-polar term. The first application of QM-PBSA with ONETEP has been on protein-protein interactions [33]. The results obtained were in good agreement with MM-PBSA, most likely because the force field employed has been extensively and carefully parameterised for protein systems and improved over a number of years. This was applied to a model of a host-guest system, [35] where the force fields are much more general and harder to parameterise. Here a significant improvement was seen over MM-PBSA for relative binding free energies. However, this does not calculate the QM solvation energy in the QM calculation, and as such is susceptible to the errors in the computed MM solvation energy.

In this work we are presenting the first QM-PBSA study of a protein-ligand system where the entire protein of 2602 atoms has been described by DFT calculations. These calculations have been performed with ONETEP, and in contrast to our previous QM-PBSA studies, the solvation free energy has been computed within ONETEP with a newly implemented self-consistent minimal parameter implicit solvation model. [10]

## 2.2 The ONETEP approach

The ONETEP [8] program is a linear-scaling DFT code that has been developed for use on parallel computers [36]. ONETEP combines linear-scaling with accuracy comparable to conventional cubic-scaling plane-wave methods, which provide an unbiased and systematically improvable approach to DFT calculations. Its novel and highly efficient algorithms allow calculations on systems containing tens of thousands of atoms [37]. ONETEP is based on a reformulation of DFT in terms of the one-particle density matrix. The density matrix in terms of Kohn-Sham orbitals is

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{n=0}^{\infty} f_n \psi_n(\mathbf{r}) \psi_n^*(\mathbf{r}'), \quad (6)$$

where  $f_n$  is the occupancy and  $\psi_n(\mathbf{r})$  are the Kohn-Sham orbitals. In ONETEP the density matrix is represented as

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha} \sum_{\beta} \phi_{\alpha}(\mathbf{r}) K^{\alpha\beta} \phi_{\beta}^*(\mathbf{r}'), \quad (7)$$

where  $\phi_\alpha(\mathbf{r})$  are localised non-orthogonal generalised Wannier functions [38] (NGWFs) and  $K^{\alpha\beta}$ , which is called the density kernel, is the representation of  $f_n$  in the duals of these functions. Linear-scaling is achieved by truncation of the density kernel, which decays exponentially for materials with a band gap, and by enforcing strict localisation of the NGWFs onto atomic regions. In ONETEP, as well as optimising the density kernel, the NGWFs are also optimised, subject to a localisation constraint. Optimising the NGWFs *in situ* allows for a minimum number of NGWFs to be used whilst still achieving plane wave accuracy. The NGWFs are expanded in a basis set of periodic sinc (psinc) functions [39], which are equivalent to a plane-wave basis as they are related by a unitary transformation. Using a plane wave basis set allows the accuracy to be improved by varying a single parameter, equivalent to the kinetic energy cut-off 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 [40].

A minimal parameter implicit solvent model has recently been developed within ONETEP [10, 41]. In this model the total potential of the solute is obtained by solving the nonhomogeneous Poisson equation within the self-consistent calculation in ONETEP:

$$\nabla \cdot (\epsilon[\rho]\nabla\phi) = -4\pi\rho_{\text{tot}}(\mathbf{r}). \quad (8)$$

Where  $\rho_{\text{tot}}(\mathbf{r})$  is the total charge density and is calculated as a sum of the electronic density  $\rho(\mathbf{r})$  and the density of the atomic cores. The solute cavity is constructed directly from the electronic density of the solute, which reduces the number of parameters required to only two [42]. The model includes a smooth transition of the relative permittivity according to the following expression:

$$\epsilon(\mathbf{r}) = 1 + \frac{\epsilon_\infty - 1}{2} \left( 1 + \frac{1 - (\rho(\mathbf{r})/\rho_0)^{2\beta}}{1 + (\rho(\mathbf{r})/\rho_0)^{2\beta}} \right), \quad (9)$$

where  $\epsilon_\infty$  is the bulk permittivity,  $\beta$  controls the smoothness of the transition of  $\epsilon(\mathbf{r})$  from 1 to  $\epsilon_\infty$ , and  $\rho_0$  is the density value for which the permittivity drops to half that of the bulk. This model was validated on two test sets, one of 60 small molecules (20 neutral, 20 cationic and 20 anionic) and the second of 71 larger molecules, on which the solvation energies obtained had a root mean square (rms) error with respect to experiment of 3.8 kcal/mol, while the PCM model [43] showed an rms error of 10.9 kcal/mol and the highly parameterised state-of-the-art SMD model [44] had an rms error of 3.4 kcal/mol.





## 2.3 Simulation details

The lysozyme structure was protonated with the MOE [45] program. MM simulations were carried out using the AMBER10 [46] package, with the ff99SB [47] force field used for the protein and the generalised AMBER force field [48] (GAFF) used to model the ligands. Ligand charges were calculated with the AM1-BCC method with antechamber (part of AMBER). The system was explicitly solvated in the TIP3P water model [49] and the charge neutralised by  $\text{Cl}^-$  ions.

Since the binding modes of the ligands are all very similar, only catechol bound in the pocket (PDB: 1XEP) was equilibrated. All other ligands were mutated from the catechol at the end point of the equilibration. The system was equilibrated using the following protocol. Hydrogen atoms were relaxed with restraints placed on all heavy atoms in the complex and solvent, before relaxing the solvent with restraints only on the complex. The system was heated to 300 K over 200 ps, still restraining the heavy atoms of the complex, in the NVT ensemble, then run for a further 200 ps with the NPT ensemble at 300 K in order to equilibrate the solvent density. This was cooled over 100 ps to 100 K and a number of relaxations were run, reducing the restraints on the heavy atoms in stages (1000, 500, 100, 50, 20, 10, 5, 2, 1, 0.5 kcal/mol $\text{\AA}^{-2}$ ). Finally the system was reheated to 300 K with no restraints over 200 ps and then for a further 200 ps at 300 K in the NPT ensemble. At this stage, it was confirmed that the water density in the box and the protein structure were stable, as measured by the root mean squared deviation of the protein backbone atoms (converged to 0.8  $\text{\AA}$  relative to the initial configuration).

Production simulations were run for 20 ns in the NVT ensemble at 300 K, with the first 1 ns being considered as further equilibration of the ligand in the pocket. All MD simulations used the Langevin thermostat [50], the particle mesh Ewald (PME) sum for the treatment of the electrostatics and the SHAKE algorithm [51] to constrain hydrogen-containing bonds, allowing a time-step of 2 fs. For the MM-PBSA calculation an infinite non-bonded cutoff was used with a dielectric constant of 80 to represent the water solvent. The 1-phenylsemicarbazide ligand was treated slightly differently. Since it is significantly bigger than catechol, this ligand was built into the protein pocket using the structure of the benzyl acetate complex with L99A/M102Q double mutant of T4 lysozyme (PDB:3HUK), which is structurally similar to 1-phenylsemicarbazide. This structure was then equilibrated in the same way as catechol, followed by a 20 ns NVT production simulation, the last 19 ns of which were used to generate the snapshots for this study.



In the ONETEP calculations, 4 NGWFs were used to describe carbon, oxygen and nitrogen atoms, 1 NGWF for hydrogen atoms and 9 NGWFs for the halogen atoms, all with radii of  $8 a_0$ . A kinetic energy cutoff of 827 eV for the psinc basis set was used, with the GGA exchange-correlation functional PBE [52] combined with our implementation of the DFT+D approach to account for dispersion, parameterised specifically for this functional [34]. These settings for our ONETEP calculations were previously compared against calculations with large near-complete Gaussian basis sets and shown to agree with them to less than 0.02 kcal/mol [53]. The numerical parameters for QM implicit solvation model were chosen after validation using this protein in a previous study [9]; these are a dielectric constant of 78.54 to represent the water solvent, a  $\beta$  value of 1.3, and numerical parameters for the smeared ion width of  $0.8 a_0$  and a discretization order of 8.

The change in entropy was computed with normal mode analysis in AMBER10. The structural optimisations were performed first with the conjugate gradients method, followed by the Newton-Raphson method, with very small force tolerances for tight convergence. These were done in the presence of an implicit solvent using the Generalised Born model.

It is important to note that the QM total energies of the complex and host are of the order of millions of kcal/mol, in contrast the binding energies are only a few tens of kcal/mol, a minute value in comparison. Thus, for the accurate calculation of the binding energies the total energies have to be very well converged, and for systems of this size (2600+ atoms) this can be challenging. We examined the convergence of the total energy on one of our large systems (phenol bound in the cavity of T4 lysozyme L99A/M102Q) to ensure that the energies are converged sufficiently well for the calculation of binding energies. The results for two snapshots of phenol bound in the pocket are displayed in Table I. For our test calculations on this system with ONETEP we see very good convergence with errors less than 0.1 kcal/mol.

Snapshots were extracted at constant time intervals from the production trajectory. For MM-PBSA the binding free energies were calculated with an increasing number of snapshots up to a total of 1000 snapshots taken from the 19 ns production simulations. Figure 1 displays the convergence of the MM-PBSA free energies as more snapshots are included in the ensemble, considering the value at 1000 snapshots as the fully converged (standard errors below 0.08 kcal/mol). The maximum error observed using 5 snapshots is 1.15 kcal/mol, for the 2-methylphenol ligand. By using 50 snapshots the maximum error is reduced to less than 0.5 kcal/mol (with catechol having the largest error of 0.41 kcal/mol). Assuming the convergence

pattern for the QM calculations is similar, and taking into account the increased computational cost of the QM calculations, in this study we have chosen to use 50 snapshots to calculate the QM-PBSA energies. We can further support the assumption of similar convergence rates if we examine the standard errors for the 50 snapshots for the MM and QM binding energies (both in solvent). For the case of phenol the MM standard error over 50 snapshots is 0.34 kcal/mol (and 0.34 kcal/mol in solvent), while the QM value is 0.35 kcal/mol (0.42 kcal/mol). For catechol the MM error is 0.34 kcal/mol (0.27 kcal/mol) compared to a QM value of 0.38 kcal/mol (0.42 kcal/mol), and for toluene the MM standard error is 0.22 kcal/mol (0.27 kcal/mol) compared to a QM value of 0.24 kcal/mol (0.34 kcal/mol). Thus, MM and QM standard error values are very similar, supporting the assumption that the QM binding energies converge at a similar rate as the MM binding energies shown in Figure 1.

Another issue for the T4 lysozyme L99A/M102Q is that of Val111 in the binding pocket, which is known to have two rotamers with a  $\chi_1$  angle of  $\sim 180^\circ$  and  $\sim -60^\circ$ . Studies have shown that using the wrong rotamer can have an effect of up to 4 kcal/mol on the calculated binding free energies [14, 19]. Explicit modelling of this effect has been shown to improve the agreement with experimental binding free energies [19]. We confirmed that our simulations sampled both of these rotamers.

When relative binding free energies are calculated, an approximation that can be used, if ligands are of a similar size and shape, is that the changes in entropy of binding between different ligands are comparable and will cancel each other. To investigate the effect of this approximation for this system, we have also approximated the entropy using normal mode analysis in AMBER and the results with and without entropy are shown. In Table III the entropy is calculated as the average from the 50 snapshots taken at constant time intervals from the trajectories.

Ligand hydration energies were calculated using the SMD model [44] in the Gaussian03 program [30] at the M05-2X/6-31G(d) SCRF(IEFPCM,Solvent=Water,SMD) level on a single geometry optimised structure.

### 3 Results and Discussion

We computed the binding free energy of the 8 ligands shown in Table II to the T4 lysozyme double mutant L99A/M102Q using MM-PBSA and QM-PBSA. These ligands were chosen as they comprise of a variety of chemical and physical properties (polarity, inclusion of halides, size and binder/non-binder).



For a protein like this with a buried binding pocket which contains solvent, both the QM and MM models will provide a qualitatively wrong description for the host by adding rather than subtracting the non-polar contribution of the cavity, as depicted in Figure 2.

We can correct the error described in Figure 2 if we subtract twice the difference between the host and complex non-polar terms from the host non-polar solvation energy; once to remove the presence of the erroneous addition of the cavity (making it equal to the description of the complex), and a second time to create a solvent filled cavity. With the observation that the surface tension term used in the QM model takes into account the solute-solvent dispersion and repulsion [10], we wish to remove only the spurious cavitation from our model of the host, whilst leaving the physically correct dispersion-repulsion contributions. The corrected host solvation energy is then given by Equation 12.

$$\begin{aligned}
 \Delta SASA &= SASA_{\text{host}} - SASA_{\text{complex}} \\
 E_{\text{cav}} &= \gamma \cdot SASA \\
 E_{\text{dis-rep}} &= -\gamma' \cdot SASA \\
 E_{\text{non-polar}} &= E_{\text{cav}} + E_{\text{disp-rep}} = (\gamma - \gamma') \cdot SASA = \tilde{\gamma} \cdot SASA. \\
 \Delta SASA &= \frac{1}{\tilde{\gamma}} \left( E_{\text{non-polar}}^{\text{host}} - E_{\text{non-polar}}^{\text{complex}} \right). \tag{10}
 \end{aligned}$$

$$\begin{aligned}
 E_{\text{cav-corr}}^{\text{QM}} &= 2\gamma \cdot \Delta SASA = 2\frac{\gamma}{\tilde{\gamma}} \left( E_{\text{non-polar}}^{\text{host}} - E_{\text{non-polar}}^{\text{complex}} \right) \\
 &\simeq 7.116 \left( E_{\text{non-polar}}^{\text{host}} - E_{\text{non-polar}}^{\text{complex}} \right). \tag{11}
 \end{aligned}$$

$$\Delta E_{\text{host-solvation}}^{\text{QM}} = E_{\text{host-solv}}^{\text{QM}} - E_{\text{cav-corr}}^{\text{QM}} - E_{\text{host-vac}}^{\text{QM}}. \tag{12}$$

Here  $\gamma$  is the solvent surface tension, while  $\gamma'$  is an additional factor to  $\gamma$  that introduces dispersion-repulsion interactions,  $SASA$  is the solvent-accessible surface area,  $E_{\text{cav}}$  is the cavitation energy,  $E_{\text{dis-rep}}$  is the solute-solvent dispersion-repulsion energy,  $E_{\text{non-polar}}^{\text{host}}$  is the total non-polar solvation energy, and  $E_{\text{cav-corr}}^{\text{QM}}$  is the QM correction to the cavitation energy. This is also the case with the MM non-polar energy, with the buried pocket adding to the surface area, causing the host to have a larger cavitation energy than the complex. We have treated the MM in a similar way, subtracting twice the difference of

the host and complex non-polar energy:

$$E_{\text{cav-corr}}^{\text{MM}} = 2 \cdot \left( E_{\text{non-polar}}^{\text{host}} - E_{\text{non-polar}}^{\text{complex}} \right).$$
$$\Delta E_{\text{host-solvation}}^{\text{MM}} = E_{\text{host-solv}}^{\text{MM}} - E_{\text{cav-corr}}^{\text{MM}} - E_{\text{host-vac}}^{\text{MM}}. \quad (13)$$

Table III presents the computed binding free energies of these ligands (relative to phenol which is normalised to the experimental binding free energy of phenol to T4 lysozyme double mutant L99A/M102Q), averaged over 50 snapshots. The top table shows the QM binding energies in vacuum, solvent, and with entropic contributions calculated from normal mode analysis (from the MM calculations), alongside the experimentally obtained values. The lower table displays the same data obtained from MM-PBSA.

Looking at the MM binding free energies (averaged over 50 snapshots) in Table III, we observe that the relative binding free energies from MM-PBSA are not very close to the experimental values. There are two exceptions, catechol that has a calculated relative binding free energy with an error of 0.7 kcal/mol when compared to the experimental value, and 2-methylphenol which has an error of 0.9 kcal/mol from experiment. The smallest error for the other binding ligands is almost 2 kcal/mol. The binding free energies from QM-PBSA are improved, with two ligands having errors less than 0.8 kcal/mol from experiment (2-fluoroaniline and toluene). Catechol, in contrast to MM-PBSA, has the largest error of the known binders with a value of 4.1 kcal/mol. Given that the vacuum binding energies from MM and QM correlate very well, as we can observe in Table III, the difference observed in the binding energies in solvent is primarily due to the solvation energies. For a deeper understanding, we will look at only the solvation energy of the ligands. We will use catechol as an example, since the MM-PBSA relative binding free energy is very close to the experimental value whilst the QM-PBSA value is not. The experimental hydration free energy of catechol is -9.4 kcal/mol [54]. The QM hydration free energy averaged over the 50 snapshots is -8.1 kcal/mol. In contrast, the MM hydration free energy averaged over the same 50 snapshots, is -20.9 kcal/mol. This shows that the MM solvation energy is substantially overestimated; however MM-PBSA still produces a very good relative binding free energy compared to experiment. Since MM and QM binding energies in vacuum are so close to each other, this suggests that both MM and QM overbind the catechol in the pocket in vacuum, however the excessively large solvation energy from MM-PBSA cancels out the overbinding in vacuum to give a final relative free energy of binding that does agree closely with experiment.

MM-PBSA predicts 1-phenylsemicarbazide, which is an experimental non-binder, to be a strong binder, while QM-PBSA correctly predicts that this molecule is a non-binder. This is a consequence of the much larger size of this ligand and the contributions this makes to the solvation energy in each model. 1-phenylsemicarbazide has the largest binding energy in vacuum for both MM and QM descriptions. Due to the larger size, 1-phenylsemicarbazide also shows more pronounced changes in the entropy of binding, which results in a difference of around 4.1 kcal/mol in the calculated binding entropy compared to phenol. Thus, the approximation of entropy cancellation is not valid in this case and indeed the inclusion of entropy has a large effect for both the MM-PBSA and QM-PBSA results for this ligand. We can assess the increase of the binding pocket volume in the case of 1-phenylsemicarbazide by examining Figure 3, which displays the dielectric permittivity from the quantum solvation model within the binding pocket at an isovalue of 70, for the host as extracted from the complex geometry with phenol bound (3.a) or with 1-phenylsemicarbazide bound (3.b). This clearly shows the enlargement of the cavity caused by the presence of the larger ligand. Solvation energies are also very important. The desolvation energy of this ligand is larger than the others due to its large polar chain and the solvation energy of the host is also increased compared to the host solvation energies for the other ligands due to the enlarged cavity, so both of these effects act to destabilise the binding of 1-phenylsemicarbazide. The combination of these effects as shown in Table III produces a positive free energy of binding showing that this ligand is a non-binder for the QM calculations, while it is still predicted to be a strong binder by the MM calculations. As the geometries are the same between the MM and QM calculations and also the relative energies in vacuum agree closely (around 3 kcal/mol), we conclude that the more rigorous QM solvation model is responsible for producing the correct result for this case. This deciding role of the solvent is also demonstrated by previous calculations using the more thermodynamically rigorous TI technique in explicit solvent in which case 1-phenylsemicarbazide was also predicted to be a non-binder [22].

It is interesting to observe that 2-aminophenol which is a known decoy [21], i.e. an experimental non-binder that computational approaches predict as a binder, lives up to its name and is predicted to be a good binder by both the MM and QM techniques in our study. By examining the binding energies in vacuum we observe that 2-aminophenol, as is the case with catechol, has very large binding energies (-12.2 kcal/mol for QM and -13.2 for MM). While this is to be expected as the amine and hydroxyl groups can each form a hydrogen bond in the cavity while most other ligands form a single hydrogen bond, the

computed binding energies appear to be larger than we might expect, even for a bidentate ligand. Graves *et al.* [21] also support this view as both catechol and 2-aminophenol are expected to share the same binding modes within the cavity. They attribute the lack of binding for 2-aminophenol to an expectation that it should have a larger desolvation energy than catechol due to the perceived ability of the amine to stabilise more hydrogen bonds. This is not confirmed however from our ligand hydration energies as calculated with three different solvation models. For example, the most accurate values we can obtain, from the QM models, for catechol and 1-aminophenol differ by 0.5 kcal/mol for the SMD model and 0.1 kcal/mol for our model. Therefore the experimental result, which has been obtained by the thermal upshift denaturation temperature technique, may need to be checked with more accurate methods. While the MM solvation energy of 2-aminophenol is overestimated by about 4 kcal/mol, this is not as large as the 12 kcal/mol overestimation of the catechol solvation energy, and does not cause the fortuitous error cancellation that led to the remarkable agreement with experiment for the MM binding free energy of catechol. If we did have the same level of error in the MM solvation energy of 2-aminophenol, MM-PBSA would predict it as essentially a non-binder.

If we exclude 1-phenylsemicarbazide for the reasons we have already discussed, all other ligands have entropies of binding much closer to phenol, between 0.4 kcal/mol and 1.6 kcal/mol, with the largest value being for 2-methylphenol. Even though this difference in entropies is quite small, we do observe an overall improvement in the agreement with experiment when entropy is included in the calculation of the relative binding free energies averaged over 50 snapshots for all ligands, for both MM-PBSA and QM-PBSA. Therefore the argument that entropies of binding for ligands of similar size can be ignored when calculating relative free energies of binding is partially valid and could be applied in cases where the calculation of vibrational entropies is not feasible. In the case of 1-phenylsemicarbazide, which is much larger and more flexible than the other ligands, the effect of the entropy is significant and cannot be ignored. A more rigorous theoretical approach for including entropy in the calculation of free energies of binding is TI. However, even with MM calculations, this approach is computationally extremely demanding compared to MM-PBSA as it requires explicit solvent and multiple MD simulations “mutating” one ligand to another. Boyce *et al.* [22] have used MM-based TI simulations to calculate free energies of binding of 8 ligands to the same protein as we. Their ligands are small rigid aromatic molecules and they have in common with us the ligands catechol, phenol and 2-methylphenol. Their rms errors of the



relative binding free energies from phenol to their other ligands with respect to experimentally measured values (obtained from ITC) were 2.5 kcal/mol, and using catechol as a reference, 1.1 kcal/mol. Our QM-PBSA calculations have comparable rms error of 2.7 kcal/mol while MM-PBSA has a larger rms error of 4.0 kcal/mol. These comparisons make it clear that the rigorous calculation of the entropy of binding is equally important to the accurate description of the intermolecular interactions that the QM method provides.

We have already discussed the effect of the solvation model, which plays a major role in the free energies of binding that each method provides. More in depth information can be obtained from Table IV where we present all the ligand hydration energies that we have available. The experimentally obtained hydration energies of catechol, 2-methylphenol, phenol and toluene are -9.4 kcal/mol [54], -5.9 kcal/mol [55], -6.6 kcal/mol [55] and -0.9 kcal/mol [55], respectively. Hydration energies obtained from QM-PBSA averaged over 50 snapshots are -8.1 kcal/mol, -2.9 kcal/mol, -3.7 kcal/mol and 1.4 kcal/mol in contrast to MM-PBSA which gives -20.9 kcal/mol, -9.0 kcal/mol, -9.8 kcal/mol and -1.5 kcal/mol. Table IV shows also the hydration energies of all the ligands as calculated with the SMD model as well. We can clearly see that the MM-PBSA hydration energies are less accurate and this impacts the outcome of the free energy calculations. The QM-PBSA energies on the other hand have a smaller error and the relative hydration energies are substantially closer to experimentally obtained values. For the classical solvation model, the maximum error for the relative solvation energies of the ligands is 8.6 kcal/mol, with an rms error 2.1 kcal/mol, compared with 1.8 kcal/mol maximum error and 1.2 kcal/mol rms error for our QM solvation model. The more accurate QM solvation energies result in improved binding free energies for the majority of the ligands.

The last thing to note is the computational effort required for these QM-PBSA calculations. As would be expected, these calculations are significantly more expensive than the MM calculations. However, it is important to remember that we are running this QM-PBSA approach on over 2600 atoms, encompassing the entire protein, and no other code is capable of running such calculations on systems of this size. A single QM-PBSA calculation on a representative complex structure took approximately 28 hours on 48 Intel® Xeon® E5-2670 processor cores. This is around  $4.8 \times 10^7$  core/seconds, compared to the MM-PBSA calculation that took 11 seconds on one Intel® Xeon® E5-2650 processor core. A difference of  $4 \times 10^6$ .



## 4 Conclusions

We have developed a QM-PBSA approach for the calculation of free energies of binding, in which large-scale DFT calculations with a near-complete basis set are performed to evaluate the energy of the configurations in place of the force field that is used in the conventional MM-PBSA technique. The solvation effects in the DFT calculations are described by a minimal parameter self-consistent implicit solvent model. We applied this QM-PBSA approach to compute the relative binding free energies of eight small aromatic ligands bound in the polar cavity of the T4 lysozyme mutant L99A/M102Q protein which contains 2601 atoms, and have compared our results to the traditional MM-PBSA method. To our knowledge, this is the first study where an entire protein-ligand system is described by a DFT approach with a self-consistent implicit solvent model.

All the structures were obtained from classical molecular dynamics simulations. The free energy calculations have been converged to within 0.5 kcal/mol with respect to the number of snapshots included in the ensembles for both the MM and QM approaches. Our aim was to explicitly account for electronic polarisation and charge transfer via the DFT calculations. We observed remarkable agreement between the MM and QM binding energies in vacuum, which indicates that the parametrisation of the force field in this case is good enough to capture, in an average way, effects of polarisation and charge transfer for the T4 lysozyme mutant L99A/M102Q. With the QM and MM vacuum energies agreeing so closely, the differences in the final binding free energies are due to the solvation energies. Thus, we observe significant differences when computing free energies of binding in solvent and we have found that our DFT-based solvation model is overall more consistent and accurate than the MM model. For the eight ligands we used in this study the rms error in free energies of binding is 4.0 kcal/mol for MM-PBSA, whereas QM-PBSA reduces this error to 2.7 kcal/mol. This demonstrates that at least the solvent-induced polarisation needs to be treated explicitly in order to improve the reliability of such free energy approaches.

Even though the QM-PBSA results are overall more accurate, the approach performs significantly worse for the catechol ligand. MM-PBSA's improved accuracy for this ligand in particular appears to be fortuitous error cancellation between overbinding in vacuum and a ligand hydration energy that is over twice the experimental value. It is important to note that the ligands in our set contain two non-binders that MM-PBSA predicts as good binders, whereas QM-PBSA correctly predicts one of these as a non-binder.

The T4 lysozyme double mutant is a challenging system, as not only do all the experimentally con-

firmed binders in our study have very similar binding free energies but also the buried cavity is a challenge for the implicit solvation models. In addition to this, the advantage of explicitly accounting for the electrons is offset by the simplicity of the protein and ligands we have used here, which, as we have shown are described very well by the force field. Although the sample size of this study was small, our full QM-PBSA approach has given very encouraging results. We expect that QM-PBSA simulations of this kind will prove beneficial in future computational drug optimisation studies, especially in cases where the ligands have functional groups whose interactions with the hosts are not captured accurately by available force fields and their parametrisations.

## 5 Acknowledgements

S.J.F. would like to thank the BBSRC and Boehringer Ingelheim for an industrial CASE studentship. J.D. acknowledges the support of the Engineering and Physical Sciences Research Council (EPSRC grant No. EP/J015059/1) and of the Polish National Science Centre and the Ministry of Science and Higher Education (grants N N519 577838 and IP2012 043972). C.-K. S. would like to thank the Royal Society for a University Research Fellowship. We would like to thank the UCKP consortium (EPSRC grant No. EP/K013556/1 ) for access to national supercomputers and the University of Southampton for the Iridis3 and Iridis4 supercomputers that were used in this work.

## References

- [1] M. K. Gilson and H.-X. Zhou, *Annu. Rev. Biophys. Biomol. Struct.* 36 (2007) 21.
- [2] B. O. Brandsdal, F. Österberg, M. Almlöf, I. Feierberg, V. B. Luzhkov and J. Åqvist, *Adv. Protein Chem.* 66 (2003) 123.
- [3] R. Zwanzig, *J. Chem. Phys.* 22 (1954) 1420.
- [4] C. D. Christ, A. E. Mark and W. F. van Gunsteren, *J. Comp. Chem.* 31 (2010) 1569.
- [5] J. G. Kirkwood, *J. Chem. Phys.* 3 (1935) 300.
- [6] I. Halperin, B. Y. Ma, H. Wolfson and R. Nussinov, *Proteins-Structure Function and Genetics*. 47 (2002) 409.
- [7] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman and D. A. Case, *J. Am. Chem. Soc.* 120 (1998) 9401.
- [8] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, *J. Chem. Phys.* 122 (2005) 084119.
- [9] J. Dziedzic, S. J. Fox, T. Fox, C. Tautermann and C.-K. Skylaris, *Int. J. Quantum Chem.* 113 (2012) 771.
- [10] J. Dziedzic, H. H. Helal, C.-K. Skylaris, A. A. Mostofi and M. C. Payne, *Europhysics Letters* 95 (2011) 43001.
- [11] B. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews and B. K. Shoichet, *J. Mol. Biol.* 322 (2002) 339.
- [12] W. A. Baase, X. J. Zhang, D. W. Heinz, M. Blaber, E. P. Baldwin, B. W. Matthews and A. E. Eriksson, *Science*. 255 (1992) 178.
- [13] A. E. Eriksson, W. A. Baase and B. W. Matthews, *J. Mol. Biol.* 229 (1993) 747.
- [14] Y. Deng and B. Roux, *J. Chem. Theory Comput.* 2 (2006) 1255.
- [15] A. Morton, W. Baase and B. W. Matthews, *Biochemistry*. 34 (1995) 8564.

- [16] A. Morton, *Biochemistry*. 34 (1995) 8576.
- [17] E. Gallicchio, M. Lapelosa and R. M. Levy, *J. Chem. Theory Comput.* 6 (2010) 2961.
- [18] M. Leitgeb, M. Karplus, S. Boresch and F. Tettinger, *J. Phys. Chem. B*. 107 (2003) 9535.
- [19] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet and K. A. Dill, *J. Mol. Biol.* 371 (2007) 1118.
- [20] Y. Deng and B. Roux, *J. Phys. Chem. B*. 113 (2009) 2234.
- [21] A. P. Graves, R. Brenk and B. K. Shoichet, *J. Med. Chem.* 48 (2005) 3714.
- [22] S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill and B. K. Shoichet, *J. Mol. Biol.* 394 (2009) 747.
- [23] A. P. Graves, D. M. Shivakumar, S. E. Boyce, M. P. Jacobson, D. A. Case and B. K. Shoichet, *J. Mol. Biol.* 377 (2008) 914.
- [24] I. Massova and P. A. Kollman, *J. Am. Chem. Soc.* 121 (1999) 8133.
- [25] M. Kaukonen, P. Söderhjelm, J. Heimdal and U. Ryde, *J. Phys. Chem. B*. 112 (2008) 12537.
- [26] T. H. Rod and U. Ryde, *Phys. Rev. Lett.* 94 (2005) 198302.
- [27] M. Wang and C. F. Wong, *J. Chem. Phys.* 126 (2007) 026101.
- [28] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejon and D. Sanchez, *J. Phys. Cond. Mat.* 14 (2002) 2745.
- [29] M. E. Davis, J. D. Madura, B. A. Luty and J. A. McCammon, *Comp. Phys. Comm.* 62 (1991) 187.
- [30] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. J. A. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts,

R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez and J. A. Pople, *Gaussian 03, Revision C.02*, (2004), Gaussian, Inc., Wallingford, CT.

- [31] N. Diaz, D. Suárez, K. M. Merz and T. L. Sordo, *J. Med. Chem.* 48 (2005) 780.
- [32] S. L. Dixon, A. van der Vaart, B. Wang, V. Gogonea, J. J. Vincent, E. N. Brothers, D. Suarez, L. M. Westerhoff and J. K. M. Merz, *Divcon*, (2004), The Pennsylvania State University, University Park, PA, 16802.
- [33] D. J. Cole, C.-K. Skylaris, E. Rajendra, A. R. Venkitaraman and M. C. Payne, *Europhysics Letters* 91 (2010) 37004.
- [34] Q. Hill and C.-K. Skylaris, *Proc. R. Soc. A.* 465 (2009) 669.
- [35] S. Fox, H. G. Wallnoefer, T. Fox, C. S. Tautermann and C.-K. Skylaris, *J. Chem. Theory Comput.* 7 (2011) 1102.
- [36] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, *Phys. Stat. Sol.* 243 (2006) 973.
- [37] N. D. M. Hine, P. D. Haynes, A. A. Mostofi, C.-K. Skylaris and M. C. Payne, *Comp. Phys. Comm.* 180 (2009) 1041.
- [38] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez and M. C. Payne, *Phys. Rev. B.* 66 (2002) 035119.
- [39] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris and M. C. Payne, *J. Chem. Phys.* 119 (2003) 8842.
- [40] P. D. Haynes, C.-K. Skylaris, A. A. Mostofi and M. C. Payne, *Chem. Phys. Lett.* 422 (2006) 345.
- [41] L. Anton, J. Dziedzic, C.-K. Skylaris and M. Probert, *Multigrid solver module for onetep, castep and other codes*, (2013), <http://www.hector.ac.uk/cse/distributedcse/reports/onetep/>.

- [42] J.-L. Fattebert and F. Gygi, *J. Comp. Chem.* 23 (2002) 662.
- [43] J. Tomasi and M. Persico, *Chem. Rev.* 94 (1994) 2027.
- [44] A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B.* 113 (2009) 6378.
- [45] *Molecular operating environment (moe), 2013.08.,* (2013.), Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7,.
- [46] D. A. Case, T. Darden, T. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, M. Crowley, S. H. 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. Mathews, M. Seetin, C., Sagui, V. Babin and P. A. Kollman, *Amber10*, (2008).
- [47] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *PROTEINS: Structure, Function, and Genetics* 65 (2006) 712.
- [48] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comp. Chem.* 25 (2004) 1157.
- [49] W. L. Jorgensen, J. Chandrasekhar and J. D. Madura, *J. Chem. Phys.* 79 (1983) 926.
- [50] S. A. Alderman and J. D. Doll, *J. Chem. Phys.* 64 (1976) 2375.
- [51] J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.* 23 (1977) 327.
- [52] J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.* 77 (1996) 3865.
- [53] S. Fox, C. Pittock, T. Fox, C. Tautermann, N. Malcolm and C.-K. Skylaris, *J. Chem. Phys.* 135 (2011) 224107.
- [54] T. Z. Mordasini and J. A. McCammon, *J. Phys. Chem. B.* 104 (2000) 360.
- [55] A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. Winget, C. J. Cramer and D. G. Truhlar, *Minnesota solvation database – version 2012, University of Minnesota, Minneapolis*, (2012).

Figure 1: Absolute deviations of the binding energies of the studied ligands as a function of the number of snapshots included in the MM-PBSA calculation, taking 1000 snapshots as the converged value.

Figure 2: Left: The host in reality. Right: The host as described by the solvation model, in terms of its contributions to the cavitation energy. In the model, the surface area of the buried pocket is added to the cavitation energy contribution of the host, providing a qualitatively wrong description of the host.  $\Delta SASA$  is the difference in the solvent accessible surface area between the complex and host.

Figure 3: The dielectric permittivity from QM calculations at an isovalue of 70 inside of the host, extracted from the complex geometry, with a. phenol bound, and b. phenylsemicarbazide bound. The ligands have been superimposed as a guide to the eye. The ligand-containing green volumes in the middle are the cavities that contribute to the solvation energy, while the apparently empty space is occupied by the protein.



Table I: Total energies and binding energies from ONETEP for two snapshots (A and B) of phenol bound in the cavity of T4 lysozyme L99A/M102Q, and the SCF convergence errors of the calculations. Energies are given in kcal/mol.

Snapshot	Complex	Receptor	Ligand	Binding energy
A	$-7360884.3 \pm 0.03$	$-7326972.6 \pm 0.03$	$-33886.6 \pm 0.000005$	-25.1
B	$-7360982.7 \pm 0.07$	$-7327066.1 \pm 0.07$	$-33886.8 \pm 0.000005$	-29.8



Table II: Ligands chosen for study in the T4 lysozyme double mutant L99A/M102Q. Experimentally measured free energies of binding ( $\Delta G_{\text{exp}}$ ) are given in kcal/mol.

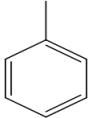
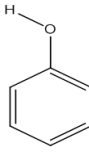
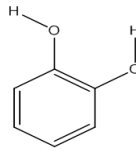
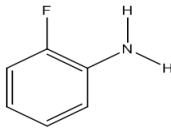
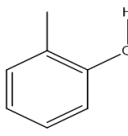
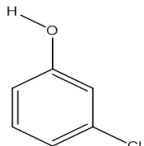
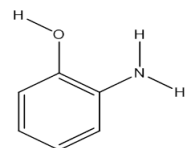
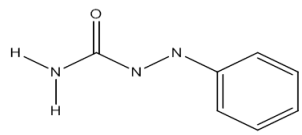
L99A/M102Q ligands	$\Delta G_{\text{exp}}$	structure
Toluene	-5.2 [11]	
Phenol	-5.5 [11]	
Catechol	-4.4 [22]	
2-fluoroaniline	-5.5 [11]	
2-methylphenol	-4.4 [11]	
3-chlorophenol	-5.8 [11]	
2-aminophenol	non-binder [21]	
1-phenylsemicarbazide	non-binder [22]	

Table III: QM-PBSA (top) and MM-PBSA (bottom) binding free energies for 50 snapshots relative to phenol. All energies in kcal/mol.

Ligand	$\Delta G_{\text{bind,vac}}^{\text{QM}}$	$\Delta G_{\text{bind,solv}}^{\text{QM}}$	$\Delta G_{\text{bind,solv}}^{\text{QM}} - T\Delta S$	$\Delta G^{\text{exp}}$
Catechol	$-13.9 \pm 0.38$	$-9.0 \pm 0.42$	$-8.6 \pm 0.92$	-4.4 [22]
3-chlorophenol	$-8.1 \pm 0.26$	$-6.9 \pm 0.40$	$-7.7 \pm 1.03$	-5.8 [11]
2-fluoroaniline	$-4.3 \pm 0.25$	$-5.9 \pm 0.42$	$-4.8 \pm 0.84$	-5.5 [11]
2-methylphenol	$-8.2 \pm 0.25$	$-8.5 \pm 0.37$	$-7.0 \pm 0.87$	-4.4 [11]
Toluene	$1.6 \pm 0.24$	$-4.8 \pm 0.35$	$-4.4 \pm 0.74$	-5.2 [11]
1-phenylsemicarbazide	$-19.8 \pm 0.49$	$-0.3 \pm 0.55$	$3.8 \pm 0.95$	non-binder [22]
2-aminophenol	$-12.2 \pm 0.41$	$-6.2 \pm 0.39$	$-5.1 \pm 0.75$	non-binder [21]
Phenol (reference)	$-5.6 \pm 0.35$	$-5.6 \pm 0.42$	$-5.6 \pm 0.81$	-5.6 [11]
Max error*	19.8	6.2	5.1	
RMS error	10.0	3.3	2.7	

Ligand	$\Delta G_{\text{bind,vac}}^{\text{MM}}$	$\Delta G_{\text{bind,solv}}^{\text{MM}}$	$\Delta G_{\text{bind,solv}}^{\text{MM}} - T\Delta S$	$\Delta G^{\text{exp}}$
Catechol	$-16.0 \pm 0.34$	$-4.1 \pm 0.27$	$-3.7 \pm 0.77$	-4.4 [22]
3-chlorophenol	$-9.6 \pm 0.25$	$-8.8 \pm 0.27$	$-9.2 \pm 0.90$	-5.8 [11]
2-fluoroaniline	$-5.7 \pm 0.28$	$-8.5 \pm 0.33$	$-7.4 \pm 0.75$	-5.5 [11]
2-methylphenol	$-8.2 \pm 0.27$	$-7.2 \pm 0.23$	$-6.1 \pm 0.73$	-4.4 [11]
Toluene	$0.8 \pm 0.22$	$-7.5 \pm 0.27$	$-7.1 \pm 0.66$	-5.2 [11]
1-phenylsemicarbazide	$-22.9 \pm 0.42$	$-11.6 \pm 0.37$	$-7.5 \pm 0.77$	non-binder [22]
2-aminophenol	$-13.2 \pm 0.40$	$-7.1 \pm 0.39$	$-6.1 \pm 0.75$	non-binder [21]
Phenol (reference)	$-5.6 \pm 0.34$	$-5.6 \pm 0.34$	$-5.6 \pm 0.73$	-5.6 [11]
Max error *	22.9	11.6	7.5	
RMS error	11.3	5.5	4.0	

\* the experimental binding energy for the non-binders is set to 0.00 kcal/mol, unless the prediction is positive, in which case the error is assumed 0.00 kcal/mol.

Table IV: Comparison of QM-PBSA and MM-PBSA ligand absolute (top) and relative (bottom) hydration energies with experimental hydration energies. Hydration energies averaged over the 50 chosen snapshots. (Energies in kcal/mol)

Molecule	$\Delta G_{\text{lig,solv}}^{\text{SMD}}$	$\Delta G_{\text{lig,solv}}^{\text{exp}}$	$\Delta G_{\text{lig,solv}}^{\text{MM}}$	$\Delta G_{\text{lig,solv}}^{\text{QM}}$
Catechol	-9.3	-9.4 [54]	-20.9	-8.1
3-chlorophenol	-6.7	-	-9.9	-3.6
2-fluoroaniline	-4.5	-	-5.4	-3.2
2-aminophenol	-9.8	-	-13.9	-8.0
2-methylphenol	-6.3	-5.9 [55]	-9.0	-2.9
1-phenylsemicarbazide	-15.1	-	-16.2	-13.8
Toluene	-1.3	-0.9 [55]	-1.4	1.4
Phenol	-6.7	-6.6 [55]	-9.7	-3.7
Relative hydration energies				
Catechol	-2.5	-2.8	-11.1	-4.3
3-chlorophenol	0.1	-	-0.1	0.1
2-fluoroaniline	2.3	-	4.4	0.6
2-aminophenol	-3.1	-	-4.1	-4.2
2-methylphenol	0.4	0.7	0.8	0.9
1-phenylsemicarbazide	-8.4	-	-6.4	-10.1
Toluene	5.4	5.7	8.3	5.1
Phenol (reference)	0.0	0.0	0.0	0.0
Max error*			8.6	1.8
RMS error*			3.4	1.2

\* Error calculation used SMD values as reference.

