Published in final edited form as:
J Phys Chem B. 2009 February 5; 113(5): 1253–1272. doi:10.1021/jp8071712.

Progresses in *Ab Initio* QM/MM Free Energy Simulations of Electrostatic Energies in Proteins: Accelerated QM/MM Studies of pK_a, Redox Reactions and Solvation Free Energies

Shina C. L. Kamerlin[1], **Maciej Haranczyk**[1,2], and **Arieh Warshel**[1]

1*Department of Chemistry, University of Southern California, 418 SGM Building, 3620 McClintock Avenue, Los Angeles, CA 90089-1062, USA*

2*Computational Research Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720-8139, USA*

Abstract

Hybrid quantum mechanical / molecular mechanical (QM/MM) approaches have been used to provide a general scheme for chemical reactions in proteins. However, such approaches still present a major challenge to computational chemists, not only because of the need for very large computer time in order to evaluate the QM energy but also because of the need for proper computational sampling. This review focuses on the sampling issue in QM/MM evaluations of electrostatic energies in proteins. We chose this example since electrostatic energies play a major role in controlling the function of proteins and are key to the structure-function correlation of biological molecules. Thus, the correct treatment of electrostatics is essential for the accurate simulation of biological systems. Although we will be presenting here different types of QM/MM calculations of electrostatic energies (and related properties), our focus will be on pK_a calculations. This reflects the fact that pK_a of ionizable groups in proteins provide one of the most direct benchmarks for the accuracy of electrostatic models of macromolecules. While pK_a calculations by semimacroscopic models have given reasonable results in many cases, existing attempts to perform pK_a calculations using QM/MM-FEP have led to large discrepancies between calculated and experimental values. In this work, we accelerate our QM/MM calculations using an updated mean charge distribution and a classical reference potential. We examine both a surface residue (Asp3) of the bovine pancreatic trypsin inhibitor, as well as a residue buried in a hydrophobic pocket (Lys102) of the T4-lysozyme mutant. We demonstrate that by using this approach, we are able to reproduce the relevant sidechain pK_as with an accuracy of 3 kcal/mol. This is well within the 7 kcal/mol energy difference observed in studies of enzymatic catalysis, and is thus sufficient accuracy to determine the main contributions to the catalytic energies of enzymes. We also provide an overall perspective of the potential of QM/MM calculations in general evaluations of electrostatic free energies, pointing out that our approach should provide a very powerful and accurate tool to predict the electrostatics of not only solution but also enzymatic reactions, as well as the solvation free energies of even larger systems, such as nucleic acid bases incorporated into DNA.

Keywords

QM/MM; Free Energy Simulations; pK_a Calculations; Solvation Free Energies; Redox Reactions

Correspondence to: Arieh Warshel, warshel@usc.edu.

I. Introduction

Hybrid quantum mechanical / molecular mechanical (QM/MM) approaches have in recent years become the key tool for calculation of protein function in general and for studies of chemical processes in proteins in particular [1–12]. In fact, well over 600 QM/MM articles were written in 2007 alone. Significant progress has been made in this direction by use of calibrated semi-empirical QM/MM approaches[1–3,11]. Some of these studies sample the phase space of the QM atoms and the surrounding MM atoms, using free energy perturbation approaches that date back to the 1980s[13]. More recent approaches that allow one to obtain the free energy surfaces for ab-initio QM/MM models[14–19] will be considered below. However, before we consider advance in QM/MM treatments it is useful to discuss some other approaches for studies of enzymatic reactions and related properties.

A seemingly obvious option for modeling enzymatic reactions is the use of high level quantum mechanical approaches. Some such gas-phase studies have been highly instrumental in providing insights about the reacting system[20] (for instance, the work of Ref.20). However, such gas-phase studies, while useful, can also be hugely problematic, in light of the fact that the environment can make a tremendous energy contribution – for instance, the solvation energies of a simple ion can be around −100 kcal/mol, and those of an infinitely spaced ion pair around −200 kcal/mol[3].

One alternative to completely disregarding the environment in the quantum mechanical calculation has been to not only consider the reacting system but to also take into account a few protein residues. While this is an improvement over examining the reacting system alone, it still cannot provide a reasonable model of the enzyme active site, particularly as it has been demonstrated that the effect of adding just a few isolated residues on the overall barrier height is negligible[21]. Also, such approaches frequently make the misassumption that residues that are ionized in the complete system would also be ionized in the model used. However, in the complete system, these ionized residues are stabilized by solvation effects of the rest of the protein, which is not the case in that particular subsystem, and thus, such residues should not be ionized[22]. A striking example of such a problematic situation is Futatsugi *et al*.'s study of the mechanism of p21ras[23], which has been discussed in detail in Refs24,[25]. The problem with this study boils down to the use of a protonated Lys16 in the reactant state, while overlooking the fact that this residue would not actually be protonated in the gas-phase environment utilized by Futatsugi *et al*. As a result, this unstable lysine ends up acting as an artificial proton relay (see the discussion in Ref.24). This is not to say that there do not exist more reasonable gas phase treatments which have paid more attention to both electroneutrality and to the system chosen. Nevertheless, when such treatments are applied to mechanistic questions, it is not impossible that they favor an incorrect mechanism, as they ignore the effect of the protein environment (for further discussion, see Ref. 26).

A related example of the problem with modeling the active site using only a few key residues is given by the work of Cavalli and Carloni[27], who came to the conclusion that Gln61 is the general base in the catalytic reaction of the Ras-GAP complex, even though it is now widely accepted that the reaction is *not* catalysed by a general base[28–30], but rather, that the key proton transfer occurs to the γ-phosphate instead[22,24,29,31]. Unfortunately, it would appear that the study of Ref.27 was performed while overlooking an extensive previous theoretical study[24] that had already demonstrated that Gln61 cannot be the general base, or of the need to validate their results by pK_a calculations. In brief, the reported study started with a subsystem that included the substrate Gln61, a few other residues and the attacking water molecule (the orientation of which was selected on the basis of already problematic force field calculations). The subsequent simulation involved an extremely short MD run while constraining the distance between the nucleophilic oxygen and the γ-phosphate to 1.8Å. This resulted in a collapse of

the water proton to Gln61, which in turn led the authors to conclude that Gln61 must be the general base. Naturally, pushing the attacking oxygen to within bonding distance from the phosphate would force the proton to be ejected to the closest base, which, in this case happened to be Gln61. However, this is neither an adequate way to examine reaction mechanisms nor to identify a general base, and a more valid study would have required not only taking into account the complete system, but also starting from a relaxed reactant state and then calculating the free energy profile for different feasible mechanisms by means of proper equilibration and long free energy perturbation simulations.

Despite this criticism, the use of a subsystem as a model for the reacting protein can actually become much more reasonable in other versions. Firstly, there exist of course clear cases where one can nevertheless obtain instructive mechanistic information from models that include a few active site residues. The most notable example of this is the case of large metal clusters, where, comparatively, the effect of the environment is relatively unimportant[32,33]. Secondly, recent studies have started to include a relatively large number of active site residues in *ab initio* studies of enzymatic reactions[34–40]. Such studies can clearly provide important mechanistic information and instructive insight about enzymatic reactions. However, we are not aware of cases in which such approaches have been able to qualitatively reproduce the catalytic effect of an enzyme (the difference between the enzymatic and the reference solution reaction). An example of this would be the case of the catalytic power of enzymes containing coenzyme B_{12} cofactor[41] (which we addressed for instance in Ref.41), which has not been yet reproduced by high level *ab initio* approaches of the cofactor and the neighboring protein residues. That is, when the topic of concern is differential effects associated with the environment (which is a central issue in biocatalysts). As another related example, a major problem would be evaluating the side chain pK_a of residue 66 in Staphylococcal nuclease (which will be discussed in Section IV). Here, high level *ab initio* approaches will most likely give very elevated pK_as, as they cannot account for protein rearrangement and water penetration upon the change in protonation state. Similarly, we believe it is unlikely that high level *ab initio* approaches can be used to deal with subtle issues like reproducing the LFER in proteins or the effects of distant mutations.

At any rate, it is important to bear in mind that outer-sphere interactions contribute in a major way to the energy changes associated with protein function. Clearly, it is not sufficient to merely treat the active site with a highly sophisticated *ab initio* method where the environment is neglected, or, at best, represented in an over simplified way as a dielectric cavity model. Additionally, a further challenge arises from the fact that proteins are not rigid but rather are highly flexible and assume many different configurations at ambient temperature. Apparently, the energy surface for different functional processes of the protein can depend strongly on protein configurations. This in turn makes it essential to average the energetics of the given process over a substantial amount of protein configurations, and the crucial need for correct sampling was already emphasized in empirical valence bond studies[42,43] in 1980. Clearly, an accurate quantum mechanical investigation in the condensed phase must reflect a compromise between the rigorous treatment of the active site and a reasonable treatment of the solvation of the active site by its protein and solvent environment. This issue becomes increasingly important the more the reaction coordinate constitutes the outer-sphere reorganisation.

A seemingly reasonable approximation to a full QM/MM treatment has been the use of *ab initio* calculations in the gas phase in order to obtain the charges and force field of the substrate (solute), and to subsequently use these charges in free energy calculations not only in solution but also in the enzyme active site[44,45]. This approach has sometimes been referred to as the QM-FE method[44,45], though this name is somewhat misleading as it gives the false impression that the QM-FE method reflects a proper FEP study of a QM surface, thus allowing for confusion with true QM/MM-FEP approaches (a more proper name would be the solvated gas

phase (SGP) model). The QM-FE approach was originally introduced by Jorgensen and coworkers[46], who applied it to studies of solution reactions. While this approach can be somewhat useful, it too suffers from some serious shortcomings. Firstly, this treatment constrains the reacting system along the gas phase reaction coordinate[44,46], which prevents one from correctly accounting for the solute entropic contributions (see below for further discussion of this issue). Secondly, this approach does not account for the fact that solute charges in the enzyme active site may be very different from those in the gas phase, and thus, this can lead to significant inaccuracies in studies of reactions where the difference between the solvation energies of the correct charges obtained in solution and those of the gas phase charges can be greater than 100 kcal/mol[47]. Despite this, the QM-FE approach still presents a major advantage over QM/MM approaches that do not perform proper configurational averaging. It should be pointed out however that all shortcomings of this approach can easily be removed by instead using an empirical valence bond (EVB) approach[16,48]. Here one also calibrates the potential surfaces and charges using gas phase information (such as the results of *ab initio* quantum mechanical calculations), but the EVB then also provides a consistent way of transferring this information to the solvating environment (see below).

It is useful to note that, despite the similar names, there is a distinct difference between the QM-FE approach as introduced by Jorgensen and coworkers and the QM/MM-FE approach of Yang *et al*[14]. The latter case is a regular QM/MM approach in which the solvent is introduced into the QM Hamiltonian (thus overcoming the fundamental inaccuracy of the approach of Jorgensen *et al.*), and the optimum path for the QM subsystem is not obtained in the gas phase but rather in the enzymatic subsystem, in the FEP calculations. That is, the approach of Yang and coworkers is basically a proper QM/MM method with incomplete FEP treatment.

Although QM/MM approaches are essential for modeling enzymatic reaction (at least until one can afford to represent all the enzyme quantum mechanically), the use of QM/MM strategy without proper sampling is not so effective[49]. This is a painful fact, despite the popularity of using such energy minimization approaches (as an example of this fact, we bring here Refs. 50–54). One option is to try to reproduce the energy minimization with many different starting points[55] (as was, for instance, done in Ref.55). However, there is no real substitute for sampling in many dimensional systems. Addressing the importance of the need for properly sampling *ab initio* QM/MM (QM(ai)/MM) surfaces has led to several important advances in this direction[14,15,19,56–62]. Many of these approaches[19,58–62] exploited the idea[15,56] of utilizing a classical potential as a reference for the QM/MM calculations, though other strategies[63–66] have also been quite promising. Also, despite some misunderstandings[66], this approach has never had convergence problems.

A key point that has perhaps been overlooked by some studies[19,58] is the non-trivial challenge of obtaining not only the free energy of the surrounding solvent, but also of the solute. This becomes crucial for solute entropy issues, as simulations using a fixed solute overlook the entropic contribution of the solute configuration which cannot easily be estimated by a harmonic approximation, at least not when one is examining the activation free energies of reactions in condensed phases[67]. Additionally, estimating the QM/MM solute path along a fixed energy minimized path is problematic in itself, as the landscape of the enzyme active site is quite complex and therefore such an approach can reflect artificial minima on the energy landscape[49]. We already demonstrated in our early works that obtaining converging results by use of a classical reference potential is quite straightforward when the solute is fixed[68]. The true problems start when the solute is allowed to fluctuate, as in such a situation, there can be a significant difference between the QM/MM potential and the reference potential. Over the past few years, we have focused significant effort into resolving this problem. This ultimately led to the development of a linear response approximation treatment by which we obtained reasonable convergence even in the quite challenging case of the autodissociation of water in

water[56]. Unfortunately, at present, the computational cost of studying chemical reactions in which the solute is allowed to fluctuate by means of QM(ai)/MM remains prohibitive, due to the requirement of very extensive configurational sampling, which, in turn, results in the extremely computationally expensive repeated evaluation of the QM energies. However, despite this, there are many cases where keeping the solute fixed could actually be desirable, for instance when studying redox reactions[57], or for pK_a calculations (which are of most relevance to the present work).

In principle, it is also possible to perform a quantum mechanical treatment of the entire protein/substrate/solvent system, and promising progresses has been made in this direction by use of a divide and conquer (D&C) approach. This approach was originally developed for *ab initio* DFT studies[69], and the principle here is to divide a large system into many smaller subsystems, separately determine the electron density of each of these subsystems, and then collect the corresponding contributions from each subsystem in order to obtain the total electron density and energy of the system. While promising, such an approach is still far too computationally expensive to be used in free energy calculations of enzymatic reactions. Thus, most current effort in treating the entire system by QM approaches has been invested into semiempirical treatments by means of various tricks to accelerate the solution of the large SCF problem. Recent efforts (which are partially inspired by the frozen-DFT method which will be discussed shortly) have resulted in a "frozen density matrix" approach[70,71], which fixes the molecular orbital coefficients of the reaction around the reacting fragments. However, again, this is not practical in terms of *ab initio* studies of enzymes, as the cost of evaluating the relevant integrals exceeds the cost of evaluating the density matrix.

Overall, the development of D&C approaches is very important as a methodological advance for computational chemistry. However, its prospects for studying enzymatic reactions are still somewhat problematic, as trying to include the entire enzyme in the QM region usually results in less emphasis on the relevant reacting fragments or on proper configurational averaging. Additionally, the effort in terms of computer time and intellectual resources in the QM description of regions far from the active site reduces focus from the relevant issues. Nevertheless, there is no question that there will be progress in representing the entire protein quantum mechanically in the future. One of the most promising options for this so far has been provided by the frozen and constraint DFT approaches (FDFT and CDFT respectively)[57,72,73]. These approaches are distinct from the frozen density matrix approach, and split the system into two regions, a region comprising the solute, any other key residues / solvent molecules (region I) and the rest of the system (region II). Here, the entire system is treated using *ab initio* DFT, but the electronic densities of the groups in region II are frozen (or constrained). The coupling between the two regions is then evaluated using a non-additive kinetic energy functional, making this approach more rigorous than the frozen density matrix approach (as we do not have to worry about orthogonality between the wave functions of the two regions since the key point is the DFT treatment of the electron density). Thus, the FDFT approach presents a way of coupling two subsystems by means of an orbital-free and first-principle-effective potential, making it possible to cast the concept of an "embedding potential" in DFT terms. This is formally related to the work of Cortona[74], though that study did not deal with the issue of embedding the subsystem into a larger system. Also, Wesolowski and Warshel[75] were the first to notice that the coupling term in any hybrid method can be obtained by partially minimizing the total energy functionals. It should be noted that even though the FDFT/CDFT approach evaluates the interactions between regions I and II quantum mechanically, it evaluates the interactions within region II classically. Thus, this approach focuses on the energy of region I, and its correct quantum mechanical interaction with region II. This allows the usage of the EVB as a reference potential for the CDFT, which means one can accurately evaluate the free energy of reactions in condensed phases and proteins while simultaneously treating the solute-solvent interaction by an *ab initio* quantum mechanical

approach[72]. The FDFT/CDFT approach has, amongst other things, been used to properly represent metal-to-ligand charge transfer in metalloenzymes[57,73] as well as studies of molecules on metal surfaces[76,77] and proton transfer reactions[78].

At this stage, it is necessary to comment on the Car-Parinelli molecular dynamics (CPMD) approach[79], which has emerged in recent years as an effective way for studying complex molecular systems, such as proton transfer reactions in solution[80]. However, using this method in gas-phase calculations of the reacting fragments with only a few protein residues has led to confusing results, such as in the case of Cavalli and Carloni's study of the Ras-GAP complex[27]. Despite this, the option of embedding the CP method in an MM surrounding and thus adopting the QM/MM philosophy[81] provided a reasonably powerful approach. This so-called Car-Parinello QM/MM approach, which has been utilized in studies of problems as varied as, for instance, transition metal catalysis[82], ion selectivity[83], polymerization reactions[84] and simulations of excited states[85]. There is no doubt that using the CP formulation as the QM part of the QM/MM approach is a reasonable strategy. However, the use of the name CP-QM/MM may give the incorrect impression that the QM/MM idea is due to the CP method and thus perhaps calling the method the QM(CP)/MM approach instead would be more appropriate.

The Car-Parinello molecular dynamics approach can be combined with the metadynamics method, in order to study processes which are dominated by changes in chemical structure, such as chemical reactions. This approach[86] (which in many respects resembles earlier ideas[87–90]) has been reviewed[91] in Ref91. Effectively, the core of the metadynamics method is to build the best reference potential, which is one that is the most similar to the actual potential (i.e. one in which the Gaussian potential approaches – E(r)). That is, repulsive markers are placed in a coarse time line in the space that is spanned by a small number of chemically relevant collective variables (CVs). These markers are then placed on top of the underlying free energy landscape in order to push the system to rapidly accumulate in the initial basin by discouraging it from revisiting points in the configurational space. In this way, the system is allowed to escape over the lowest transition state as soon as the growing biasing potential and the underlying free energy well exactly counterbalance each other, effectively allowing the simulation to escape free energy minima.

As with the combined Car-Parrinello QM/MM approach, the metadynamics approach has rocketed in popularity in recent years and has also been applied to a wide range of problems, such as flexible docking in solution[92], photoinduced ring transformations[93], intramolecular reactions in complex systems[94], phase transitions[95], rational catalyst design[96] and organometallic reactivity[97]. However, it should be noted that although the metadynamics approach has been formulaetd in an elegant way, the impression of having a new powerful way to solve the sampling challenge is problematic and counterproductive as far as finding effective approaches is concerned. That is, the philosophy of this approach is actually almost identical to the overall philosophy of our earlier approach of using a reference potential. In fact, the construction of a potential that makes the landscape flat is similar to the use of a simplified folding model as a reference potential[88], and the general use of a reference potential for accelerated sampling has been the key part of our QM/MM-FEP studies for a long time[15]. Futhermore, while it is currently very fashionable to use approaches where the best reaction coordinate is not assumed *a priori*[98–100], we believe that using chemical knowledge can in many cases be superior to a blind search (even though many workers prefer black box approaches). More specifically, using for instance the EVB as a reference potential includes an enormous amount of chemical information, which is clearly an advantage and not a disadvantage. At any rate, we are not aware of any metadynamics calculations that have provided reliable *ab initio* QM/MM free energy surfaces with less effort than that required for our reference potential approach.

Since metadynamics is basically an approach that has been geared towards getting the best reference potential, we can (with the aim of being instructive rather than humorous) define a "paradynamics" method that starts by evaluating the real potential on a rough grid of nxm points (this search can be done by very short MD simulations with a constraint on each grid point, or, even better, by evaluating $E(r_{nm})$). Then, we can fit an EVB potential to the grid, run FEP on the EVB, and, finally, evaluate the corrections due to the difference between E_{EVB} and E_{real}. Of course, we can further refine the EVB surface and minimize $<E_{EVB} - E_{real}>$ by optimizing the EVB parameters. We believe that this paradynamics approach will be faster and more accurate than the current implementation of metadynamics.

Although this review focuses on *ab initio* QM/MM approaches, we would like to emphasize the fact that in many cases, one can obtain major insights into processes in proteins by first calibrating potential surfaces to both experimental and *ab initio* information about the relevant solution reaction, and then moving the calibrated surface to the protein active site. This philosophy was introduced by Warshel and coworkers starting in 1980[42], and has been applied to a wide range of functional issues such as pK_a calculations, redox reactions, enzymatic reactions, and so forth. This has best been formulated by the EVB method[16,42,48,101]. The EVB describes the reacting system by diabatic states which include the effect of environment and then mix these states to obtain the actual potential surface. The mixing terms and the diabatic energies are calibrated to experimental and *ab initio* information about the gas phase and solution surfaces. These are then transferred to the protein while keeping all the parameters unchanged and only replacing the interaction with water in the reference reaction with the interaction with the protein/solvent in the active site. It is important to note that after some initial criticism, the EVB has become a major simulation tool, with more and more research groups taking it or closely related approaches as a powerful modeling approach[48,102]. In fact, many workers have adopted more or less all the key elements of the EVB approach, though they have used somewhat different names[101,103–105] (as has been discussed in Ref.103).

In view of the proliferation of the use of the EVB, and its reincarnation as seemingly different approaches, it is instructive to discuss the recent attempts of Truhlar and coworkers[105–107] to capture the physics of the EVB approach under a new name. This was initially started with gas-phase studies[106,107] under the name "Multi Configurational Molecular Mechanics" (MCMM), which is effectively an identical approach to the EVB, as has already been discussed in Refs.101,[103]. The more recent attempt to extend the EVB to studies in solution[105] under the name "electrostatically embedded MCMM based on the combined density functional and molecular mechanical method" is more problematic, since to a superficial reader, it may appear to be a novel and effective innovation. However, the electrostatically embedded MCMM (which we refer to here as the EE-MCMM(EVB)) is basically identical to regular EVB, with the exception of one minor modification that is in itself problematic (except for cases where it has negligible effect). That is, this method has the same diagonal EVB elements and the *same* crucial embedding by the interaction of the diabatic charges with the solvent potential as in the original EVB. Thus, the solvent is incorporated into the EVB diagonal elements, as is done by the standard EVB treatment. Even the addition of the solute polarization in the EVB states has long been implemented in some EVB studies.

In the EE-MCMM(EVB) approach, the gas phase off-diagonal (H_{12}) term is evaluated in the same way as is done in the EVB approach. However, the attempt to evaluate this term in solution is problematic, except when the solvent has no appreciable effect on H_{12}, as is the case with the EVB treatment (see below). Also, the solute diabatic charges are evaluated by DFT in the gas-phase exactly as in the EVB, and thus, the impression of coupling to DFT may be misleading, as it gives the impression that this is a novel QM/MM approach (the issue of H_{12} is dealt with separately below). The implementation in Ref.105 evaluates the potential from the solvent by an integral equation rather than by a microscopic MM treatment, but,

obviously, the application of this approach to enzymes will require moving back to the microscopic electrostatic treatment of the EVB.

Apparently, the only different element in the EE-MCMM(EVB) approach is the attempt to make the H_{12} term solvent dependent. This is a quite problematic endeavor for several reasons. Firstly, H_{12} is almost solvent independent as was initially assumed in the EVB and established in our recent studies[108] and those of others[109]. In fact, the result reported in Fig. 9 of Ref. 105 shows a solvent independent H_{12} in solution using the solvated diagonal elements and the expansion of V_g by applying the EVB relationship ($H_{12}=[V_{11}-V_g)(V_{22}-V_g)]^{1/2}$). This approach only gives reasonable results because the use of a solvated H_{ii} makes H_{12} solvent independent. In fact, the ground state surface V_g used to evaluate H_{12} would strongly depend on the solvent in challenging charge separation reactions, which are very different from the trivial S_N2 case studied in Ref.105, and this dependence could not be represented by a simple expansion (see below). However, once the basic and crucial embedding idea of the EVB is adopted, the off-diagonal term becomes more or less solvent independent, and any treatment that tries to asses its minor solvent dependence will look reasonable.

To make the above discussion even clearer, we would like to point out that if the expansion treatment were able to reproduce a correct description of the ground state adiabatic surface in solution, there would be no need for any EVB treatment or any diabatic V_{ii}, and the expansion would have provided the long awaited solution to the general QM/MM-FEP problem by gas-phase expansion so that no one would need an EVB type formulation. As to the problems with evaluating V_g (and thus H_{12}) by the expansion approach, we can return to our standard example of S_N1 reactions[47,48]. In this case, the gas phase system in the large separation range is a bi-radical, with zero charge on the separated atoms. The first term in the expansion will be zero since this term is the gas phase charge, thus, the expansion will give zero solvent effect on V_g (in contrast to the enormous effect obtained with the correct solvation treatment). In fact, the success of the approach of Ref.105 in the case of S_N2 reactions is to be expected, since even full gas phase charge distribution (in the QM-FE treatments) gives reasonable results[46]. However, the same results would be obtained with the EVB and constant H_{12}, and this type of EVB also works extremely well in the case of S_N2 reactions[48,110]. In summary, the EE-MCMM method is practically identical to the EVB approach, and it is a full QM/MM approach (once H_{12} is taken as solvent independent). In this case, the consistent embedding is entirely due to the effect of the solvent on the diagonal EVB elements.

This review will only briefly mention *ab initio* QM/MM-FEP calculations of chemical reactions and will focus on one important issue, namely the evaluation of electrostatic free energy in proteins. We start by clarifying that electrostatic effects reflect properties whose origin is entirely electrostatic (e.g. redox potentials, pK_a, conduction of ions in ion channels[111–121] and solvation energies of ions[122,123]), as well as properties whose origin turn out to be largely due to electrostatic effects (e.g. enzyme catalysis[22,24,124–126], proton transport, electron transport[127], protein-protein interactions[128–132], and molecular motors[125]). For both classes it is crucial to have the ability to accurately evaluate the corresponding electrostatic energy if we are interested in any quantitative structure function correlation in proteins[104,133–143]. Here, we will focus only on the first class and even in this case we will mainly consider solvation and pK_a calculations by *ab initio* QM/MM-FEP approaches. As a background, we will consider below general studies of pK_as in proteins.

Attempts to study the electrostatic properties of proteins commenced far prior to the emergence of structural information – the first theoretical approach to the problem was initiated as early as 1924 by Linderstrom-Lang[144], who evaluated the Coulombic energy of charged proteins by assuming all charges to be uniformly distributed over the surface of a spherical model protein. The availability of protein X-ray structures allowed one to move towards a more

realistic description of protein electrostatic energies. Nevertheless, this still required overcoming the challenges of providing a proper description of the complete protein-solvent system, and initial attempts to consider the intra-protein field[145,146] only considered the effect of the protein residual charges, thus overlooking the enormous dielectric effect of the surrounding solvent as well as the protein induced dipoles. Early attempts to use macroscopic formulations in the evaluation of electrostatic energies in proteins with known structures emerged with the influential Tanford-Kirkwood (TK) model[147,148]. However, this model implicitly assumes that all ionisable groups are on the protein surface, thus overlooking the self energy term[149] (for more details see the discussion in Ref.149). The microscopic Protein Dipoles-Langevin Dipoles (PDLD) model[4,136,150,151] has identified and overcome the fundamental problems associated with the self-energy and the corresponding intrinsic pK_a and has led to the realization that the local polar environment (which has not been considered in the early macroscopic models) plays a crucial role in determining the energetics of ionized residues. The PDLD approach provided the first physically consistent treatment of electrostatic effects in proteins by simultaneously representing the microscopic dielectric of the protein and the surrounding water molecules, while overcoming the problems and fundamental uncertainties associated with the macroscopic models (by treating the solvent molecules explicitly), as well as some of the convergence problems associated with all-atom solvent models (by treating the average solvent polarization rather than averaging the actual polarization energy)[152]. The PDLD model facilitated the early consistent treatments of pK_as of ionisable residues in proteins[4,134,151,152]. These studies were then subsequently augmented by studies using the semi-macroscopic version of the PDLD model (the PDLD/S-LRA model)[153–155].

Macroscopic treatments of the pK_a in proteins progressed from the Tanford-Kirkwood (TK) model[147,148] to discretised continuum treatments[156–158], but have not originally included the effect of the protein permanent dipoles[159] (for further details see discussion in Ref.159). These models have eventually evolved to a physically consistent description of the effect of the surrounding solvent (by using a numerical Poisson-Boltzmann treatment), as well as to the gradual incorporation of the effect of the protein permanent dipoles[134,149,160–176], and, in some cases, elements of the LRA treatment[169]. These models used a protein dielectric constant, whose true nature has been discussed elsewhere[133,177]. Although such macroscopic models are generally effective and widely used, they are nevertheless not the subject of the present work.

Microscopic all-atom calculations of pK_as in proteins emerged in the mid-80s with early free energy perturbation (FEP) calculations[178], and were followed by subsequent FEP[179] and linear response approximation (LRA)[153,155,178,180,181] calculations. Such studies are still relatively rare, but becoming gradually more common[182–184]. Although all atom FEP calculations involve major convergence problems[133], they provide, in principle, the most reliable classical option.

Obviously, classical FEP calculations are very useful, but here we will be mainly focusing on a single issue, namely pK_a calculations using combined QM/MM free energy perturbation (QM/MM-FEP). In this case, the main issue is the validation of the QM/MM approach rather than the determination of the specific pK_a. Recently, significant progress has been made in this area[185,186]. However, while promising, such QM/MM-FEP has in some cases still shown large deviations between experimental and calculated pK_as[185,186]. Furthermore, the above studies have involved semi-empirical QM approaches, while the true challenge is in the accurate use of an *ab initio* QM approach (QM(ai)/MM) in QM/MM-FEP calculations of pK_as in proteins, particularly as such studies have been demonstrated to provide “chemical accuracy” in gas phase studies of the reactions of small molecules[187]. To date, QM(ai)/MM calculations have proven to be particularly problematic for computational chemists, due not

only to the need for proper long-range treatments and effective boundary conditions for the protein and solvent environments[179], but, most importantly, due to the need for performing sufficient sampling in order to obtain accurate free energies[3,49,56,188], which results in extremely long computational times.

Recently, we introduced an advance in QM/MM-FEP calculations of solvation free energies[62], where the solvent environment is represented by an average solvent potential which is then added to the solute Hamiltonian (making it effectively a mean field approximation). Averaging potentials have already been implicitly implemented in both the QM/Langevin Dipole (QM/LD) model[48,89,189] and various continuum models[190–193]. The issue of using an average potential has also been taken up by Aguilar *et al.*[194–197] as well as by Yang *et al*[198]. Our new treatment uses a seemingly simple approach that maps the effect of the fluctuating MM environment over *m* time steps on a set of *m* × L point charges (where L is the number of atoms that are situated within a pre-defined and relatively small cutoff radius), while the rest of the environment is represented by a dipole. A key element of our treatment has been the use of a classical reference potential with fixed charges (see the Methodology section for further details) as a reference for the QM/MM calculations. Combining the average potential and the reference potential has been demonstrated to lead to computational time savings of up to 1000x in QM(ai)/MM-FEP calculations of solvation free energies of simple systems where the solute structure is kept fixed during the simulation[62]. Here, we will apply the same approach to evaluate the pK_as of a simple system, namely the Asp3 side chain of the bovine pancreatic trypsin inhibitor, as well as of that of Lys102 in the M102K T4-lysozyme mutant, as the evaluation of the pK_as of all ionisable groups in lysozyme has become a benchmark for testing the validity of pK_a calculations in general[155,181,185,199–201]. In both cases, we obtain results within a range of 2.5 pK_a units. By reproducing pK_a shifts of protein side chains with reasonable accuracy based on known experimental values as well as the usage of the appropriate thermodynamic cycle, we demonstrate that our approach can be considered reliable for reproducing electrostatics in enzymatic systems.

II. Outlining the Accelerated QM(ai)/MM-FEP Approach for the Calculation of Electrostatic Energies

II.1. Solvation free energies in water and proteins

As will be discussed in Section II.2. pK_a calculations require the evaluation of the solvation free energies of the neutral and ionized species in both water and protein. These can be accurately predicted by the use of our accelerated QM/MM approach, where the system is divided into two regions: of main interest to us is the part of the system which changes (e.g. by protonation or deprotonation), and this part of the system is designated as the solute. The remainder of the system is designated as the solvent. The selected regions are then described by QM and MM approaches respectively. This methodology has been presented and thoroughly discussed in Ref.62, and its application towards predicting solvation free energies of small biomolecules in water[202] was demonstrated in Ref.202. In the current study, we have made some minor changes to the protocol outlined in these works, and therefore we will briefly present our methodology.

In summary, our starting point for the evaluation of the relevant solvation free energy is by use of the free energy perturbation (FEP) adiabatic charging (AC) approach[48,178], which uses a mapping potential of the form:

$$E_k = E_{tot}(1 - \lambda_k) + E\prime\lambda_k \quad (1)$$

Here, E_{tot} denotes the total energy of the system, E' denotes the energy of the same system without electrostatic solute-solvent interactions, and λ_k is progressively modified from 0 to 1 in n+1 steps. From here, we can use the standard FEP equation[203]:

$$\Delta\Delta G_{sol}(\lambda_k \rightarrow \lambda_{k+1}) = -\beta^{-1} \ln\langle \exp\{-(E_{k+1} - E_k)\beta\}\rangle_{E_k}$$
$$\Delta G_{sol} = \sum_{k=1}^{n+1} \Delta\Delta G_{sol}(\lambda_k \rightarrow \lambda_{k+1}) \quad (2)$$

In Eq. 2, $\beta = 1/(k_B T)$, where k_B is Boltzmann's constant and T denotes the absolute temperature, $\langle\rangle_{E_k}$ represents the average obtained during the propagation of configurations that use E_k. The process can also be effectively approximated using a linear response approximation (LRA) treatment[204]:

$$\Delta G_{sol} \cong \langle E_{tot} - E'\rangle_{E_{tot}} + \langle E_{tot} - E'\rangle_{E'} + \Delta G_{cav} \quad (3)$$

where ΔG_{cav} is the solvation free energy of the non-polar neutral form of the solute. This term consists of two parts that describe the hydrophobic and van der Waals free energy of the cavity, respectively. These have been implemented in the ChemSol[189] program and are described in detail in Ref.189.

Here, we combine the FEP-AC and LRA approaches of Eq. 2 and Eq. 3 into a more efficient protocol, as was introduced in Ref.62, and presented in its final form in Ref.202. This protocol follows the cycle shown in Fig. 1, where the free energy of charging the solvated solute to a given charge distribution (which represents the solute polarized by the solvent partial charges obtained with the given implicit solvation model (ISM), which may be COSMO[205,206], the polarized continuum model (PCM)[207–211] or the Langevin Dipoles model[212]) is obtained by the FEP-AC approach, and then the free energy change associated with allowing these charges to "equilibrate" with the solvent potential is added as a correction. Q_{eq} denotes the vector of equilibrated QM/MM residual atomic charges of the solute, and Q_g and Q_{COSMO} denote gas-phase charges and charges obtained using the given ISM respectively. In all cases, the subscripts g and s denote gas-phase and solution states respectively, and the superscript 0 designates a constant value to the charges. Thus, from the cycle in Fig. 1, the QM/MM solvation free energy can be written as:

$$\Delta G_{sol} = \Delta E^{pol}_{QM}(Q^0_g \rightarrow Q^0_{ISM}) + \Delta G_{sol}(Q^0_{ISM}) + \Delta\Delta G_{sol}(Q^0_{ISM} \rightarrow Q_{eq}) \quad (4)$$

$\Delta E^{pol}_{QM}(Q^0_g \rightarrow Q^0_{ISM})$ is the solute polarization energy predicted by the COSMO solvation model. Of course, this charge set can be replaced by any other fixed set obtained from alternative implicit solvation models such as the polarised continuum model (PCM)[207–211] or the Langevin Dipoles model[212]. $\Delta G_{sol}(Q^0_{ISM})$ in Eq. 4 represents the solvation free energy of the solute, the atomic charges of which are obtained from the COSMO solvation model (not to be confused with the COSMO solvation energies). This value can be obtained by the classical FEP-AC approach using the following substitution:

$$\Delta G_{sol}(Q^0_{ISM}) = \Delta G_{sol}(Q=0 \rightarrow Q^0_{ISM}) + \Delta G_{cav} \quad (5)$$

The last term of Eq. 4 (i.e. the free energy change associated with allowing partial charges obtained by a given ISM such as COSMO to "equilibrate") can be obtained using the LRA approach. That is:

$$\Delta\Delta G_{\mathrm{sol}}(Q^0_{\mathrm{ISM}} \rightarrow Q_{\mathrm{eq}}) \cong \frac{1}{2}[<E_{\mathrm{tot}}(Q) - E_{\mathrm{tot}}(Q^0_{\mathrm{ISM}})>_{E(Q)} + <E_{\mathrm{tot}}(Q) - E_{\mathrm{tot}}(Q^0_{\mathrm{ISM}})>_{E(Q^0_{\mathrm{ISM}})}] \quad (6)$$

Here, E_{tot} (Q) is the QM/MM surface with fluctuating charge. For both terms in Eq. 6, both E_{tot} (Q) and $E_{\mathrm{tot}}(Q^0_{\mathrm{ISM}})$ are evaluated using the exact same solvent coordinates as both are propagated using the same potential. The only difference between the two systems is found in the set of residual charges of the solute. Thus, each of the average (< >) terms becomes the difference in the polarisation and interaction energies of both systems. Calculating each averaged term of Eq. 6 requires performing QM calculations as the configurations are propagated to obtain solute polarisation energies that include E_{tot} (Q). Therefore, unsurprisingly, calculating Eq. 6 is quite expensive. At any rate, using Eq. 4 to Eq. 6 we get:

$$\Delta G_{\mathrm{sol}} = \Delta E^{\mathrm{pol}}_{\mathrm{QM}}(Q^0_g \rightarrow Q^0_{\mathrm{ISM}}) + \Delta G_{\mathrm{sol}}(Q=0 \rightarrow Q^0_{\mathrm{ISM}}) + \Delta G_{\mathrm{cav}} + \Delta\Delta G_{\mathrm{sol}}(Q^0_{\mathrm{ISM}} \rightarrow Q_{\mathrm{eq}}) \quad (7)$$

Though the main time consuming step when solving Eq. 7 is still the evaluation of the LRA ($<>_{E(Q)}$) term.

For the purpose of calculating the $<>_{E(Q)}$ term, the system is represented using a QM/MM approach, where the residue to be ionised (so far referred to as the solute) is represented using QM. The total energy for such a system, E_{tot}, which depends on the solute and solvent coordinates (R and r respectively) can be approximated with:

$$E_{\mathrm{tot}} \cong E^g_{\mathrm{QM}}(R) + E^{el}_{\mathrm{QM/MM}}(R, r, \overline{Q}) + E_{\mathrm{VdW}} + E_{\mathrm{MM}} + \Delta E^{\mathrm{pol}}_{\mathrm{QM}}(Q^0_g \rightarrow \overline{Q}) \quad (8)$$

In this equation, $E^g_{\mathrm{QM}}(R)$ is the solute energy in the gas phase (which is obtained by *ab initio* QM), $\overline{Q}$ is some estimate of the solute partial atomic charges in solution and therefore it represents the solute polarization. The solute-solvent electrostatic interaction, $E^{el}_{\mathrm{QM/MM}}$, is obtained by its classical approximation (given in kcal/mol):

$$(E^{el}_{\mathrm{QM/MM}})_{cl} \equiv 332 \sum_{i(S)} \sum_{j(s)} \frac{Q^i q^j}{r_{\mathrm{ij}}} \quad (9)$$

Here, i and j are indexes for the solute and solvent atoms respectively, and Q and q represent the solute and solvent residual charges, respectively. E_{VdW} is the solute-solvent van der Waals interaction energy and the E_{MM} is the energy of solvent described with a classical force-field. Finally, $\Delta E^{\mathrm{pol}}_{\mathrm{QM}}$ is the polarisation energy of the solute, which can be represented as:

$$\Delta E^{\mathrm{pol}}_{\mathrm{QM}} = <\Psi^s|H^g|\Psi^s> - <\Psi^s|H^g|\Psi^s> \quad (10)$$

where Ψ^s and Ψ^g are the solute wavefunction in the gas-phase and in solution respectively, and H^g is the solute Hamiltonian.

The polarisation energy can be estimated either by LD, PCM or COSMO calculations, or by a few QM/MM calculations at some solvent configurations (here we used the COSMO model). It can also be evaluated by the classical approximations discussed in Ref.190. However, such approximations prove to be problematic, as the solute wavefunction fluctuates during the solvent fluctuations. Here, we instead use the averaged potential from the solvent to evaluate the polarisation by adding the solvent averaged potential to the solute Hamiltonian. Our approach (which was outlined in Ref.62) constrains the atoms of the QM region, evaluates the QM charges of all atoms in this region ($Q^{(1)}$, where (1) represents the first step of an MM/MD run) and then proceeds to run *m* MM/MD steps. All this time, we allow the solvent atoms to move in the potential ($E^{\mathrm{el}}_{\mathrm{QM/MM}}(Q^{(1)})+E_{\mathrm{VdW}}+E_{\mathrm{MM}}$). In this way, we obtain *m* snapshots of the solvent coordinates from *m* MM/MD steps. At this point, we scale the charge of each solvent atom by 1/*m*, and then send $m \times N$ solvent atoms with the scaled solvent charges to the QM program. The main problem with this approach, however, is that it can be computationally extremely intensive, as it generates $m \times N$ external charges that have to be included into the Hamiltonian within the QM program, which is not computationally feasible when dealing with very large systems. Thus, we use the approximation introduced by Ref.62 and shown in Fig. 2.

In the treatment shown in Fig. 2, the solvent has been divided into two regions. In the first region (Region I), the N_{ext} solvent charges are converted to $m \times N_{\mathrm{ext}}$ external charges (which are all scaled by 1/*m*), whereas Region II represents the average solvent field coming from $N - N_{\mathrm{ext}}$ solvent molecules by two point charges (q and q-) using the following relationship:

$$E_O = \frac{2q}{|r_{\mathrm{OR}}|^3} r_{\mathrm{OR}} \tag{11}$$

Where E_O is the electric field at point O (the geometrical centre of the QM system) and r_{OR} is a vector pointing along E_O to charge q. This approach has been extensively validated in Ref. 62.

II.2. pK_a Calculations

Evaluating the energetics of a single charged group in a protein is possibly the most unique test of the efficacy of models of electrostatic calculations in proteins - in fact, it has been argued that the ability to accurately predict enzyme rate constants is limited by the accuracy of the corresponding electrostatic calculations and thus by the accuracy of pK_a calculations[151].

Fig. 3 illustrates a thermodynamic cycle by which the free energy of ionizing an acid in a protein can be converted to the relevant pK_a values[151,152,155]. Here, the ionisation process in the protein has been represented in terms of the energy of the corresponding reaction in water, as well as the solvation of the ionised (A^-) and neutral (AH) species in the protein relative to in water. In the first step, 1 mol of neutral acid (AH) bound to the protein is transferred to a solution at a pH that corresponds to a hydrogen concentration of C_0. The energetics of this process can be given by the difference between the solvation energies of AH in water and in protein, (that is, $\Delta G^{p\rightarrow w}_{\mathrm{solv}}(AH)$). Subsequently, 1 mol of AH is ionised to give A^- in solution (while the pH remains constant). The energetics of this step is in turn obtained from the difference between the pK_a of the acidic group in water ((pK_a^w)), which is equivalent to $\Delta G_w = 2.3\mathrm{RT}(pK_a^w - pH)$. Here, ΔG_w represents the free energy of ionising the acid in water, and R is the ideal gas constant. Finally, in the last step, the solvated ionised species (A^-) is moved from water to the protein, where, once again, the energetics of this process is equivalent to the difference in the solvation energies of the ionised species in water and in the protein

$((\Delta G_{solv}^{w\to p}+(A^-))$). Thus, the free energy of ionising an acid in a protein at any given pH (ΔG_p) value[152] can be obtained by:

$$\Delta G_p(\mathrm{AH}_p \rightarrow A_p^- + H_w^+) = \Delta G_{sol}^{w\to p}(A^-) - \Delta G_{sol}^{w\to p}(\mathrm{AH}) + \Delta G_w(\mathrm{AH}_w \rightarrow A_w^- + H^+) \quad (12)$$

where *p* and *w* designate protein and water respectively, and $\Delta G_{\mathrm{sol}}^{w\to p}$ designates the free-energy difference of moving the indicated group to its protein site from water. Eq. 12 can then be further simplified[151,177], for the *i*th ionisable residue, to give:

$$pK_{a,i}^p = pK_{a,i}^w - \frac{\bar{q}_i}{2.3RT}\Delta\Delta G_{\mathrm{sol}}^{w\to p}(\mathrm{AH}_i \rightarrow A_i^-) \quad (13)$$

Here, $\bar{q}_i$ represents the charge of the ionised form of the relevant residue (which is −1 for acids and +1 for bases) and $\Delta\Delta G_{\mathrm{sol}}^{w\to p}(\mathrm{AH}_i \rightarrow A_i^-)$ consists of the first two terms of Eq. 12. The full derivation of Eq. 13 has been discussed in detail elsewhere.[151,177]

In general, it is useful to evaluate the free energy of an ionised group by first considering the self-energy of ionising this group when all other ionisable groups are uncharged, and then considering the effect of charging all other groups to their given ionisation state. Thus, we can express the pK_a of each group of the protein by:

$$pK_{a,i}^p = pK_{\mathrm{app},i}^p = pK_{\mathrm{int},i}^p + \Delta pK_{a,i}^{\mathrm{ch\ arg\ es}} \quad (14)$$

In this equation, $pK_{\mathrm{int,i}}$ is the pK_a that the *i*th group in the protein would have if all other groups were neutral (i.e. the "intrinsic" pKa), $\Delta pK_{a,i}^{\mathrm{ch\ arg\ es}}$ represents the effects of charging all other ionisable groups to their ionised state, and $pK_{app,i}$ represents the apparent (actual) pK_a of the *i*th ionisable group. Based on this, Equation 13 can now be re-written to give:

$$pK_{\mathrm{app},i}^p = pK_{a,i}^w - \frac{\bar{q}_i}{2.3RT}\Delta\Delta G_{\mathrm{self},i}^{w\to p} + \Delta pK_{a,i}^{\mathrm{ch\ arg\ es}} \quad (15)$$

Where $\Delta\Delta G_{self}$ is the self-energy associated with charging the *i*th group in its specific environment. This equation has been discussed in great detail elsewhere[151,155,177]. In this work, we followed our general strategy of evaluating the electrostatic energies in two cycles, where we first evaluate the self energy when all ionisable residues (with the exception in some cases of very close neighbors) are kept neutral, and then continue the thermodynamic cycle by evaluating the effect of the ionisable residues macroscopically[155], using a distance dependent dielectric constant (see e.g. Ref.155 for further details). Here, we return to the same approach and evaluate $\Delta pK_a^{\mathrm{ch\ arg\ es}}$ classically. We must point out in this respect that we do not find any compelling reason to evaluate this contribution quantum mechanically, since it mainly depends on the compensating dielectric that represents the protein and water reorganisation energies, which should be evaluated classically or represented by an effective dielectric constant.

Here, we demonstrate the use of the methodology presented in Sections II.1. and II.2. for evaluating the free energy of solvation of the sidechains of Asp3 in the bovine pancreatic trypsin inhibitor (BPTI) and Lys102 in the T4 lysozyme (with an M102K mutation) relative to acetate and methylamine in solution respectively for both neutral and charged species. Both acetate and methylamine were solvated by a sphere of explicit water molecules with a radius of 16Å

(Fig. 2, Region I), as well as two charges (a dipole) representing the rest of the system. In the case of the protein, however, Region I was defined as not only all water molecules but also all electroneutral groups within a radius of 10Å of the relevant sidechain (this amounted to 16 electroneutral groups for BPTI and 34 for lysozyme, designated as "P" in Fig. 2b). The smaller Region I radius in the protein in comparison to in water was necessary due to the much larger overall system size (this should not be confused with the size of the system in the classical AC simulation). Table 1 shows the number of explicit water molecules solvating different systems, as well as radii for Regions I and II for all systems studied. All solvent molecules have been represented by the ENZYMIX force field[153]. In our simulation model, the sphere of explicit water molecules is surrounded by a surface region, the average polarisation and radial distribution of which are determined by the surface-constrained all-atom solvent (SCAAS) model[134,178,213]. This surface region is embedded in a bulk continuum region with ε=80 (the appropriate value of ε depends on the electrostatic treatment and should be determined by using benchmarks that can, in addition, weed out unreliable treatments of electrostatics[177]). Finally, the long-range interactions are treated by the local reaction field (LRF) approach[179].

All MD simulations presented here were performed using the MOLARIS simulation package[153]. The classical adiabatic charging FEP calculations were performed in 26 steps of 10 (model compound in water) or 50ps (protein sidechain) each, in both the forward ($\Delta G_{sol}(Q{=}0 \rightarrow Q^0_{COSMO})$) and backward (($\Delta G_{sol}(Q^0_{COSMO} \rightarrow Q{=}0)$)) directions. The average of the forward and backward charging processes was then used as the appropriate value for $\Delta G_{sol}(Q{=}0 \rightarrow Q^0_{COSMO})$ in the cycle shown in Fig. 1.

The remaining terms from Eq. 2 were obtained using combined QM/MM calculations.

In each case, we first relaxed the system in either a 25 (solution) or 50ps (protein) long simulation, using 1 fs time steps and then ran either 250ps (model compounds in solution), 500ps (Asp3 sidechain of the bovine pancreatic trypsin inhibitor) or 1ns (Lys102 sidechain of the M102K T4-lysozyme mutant) simulations to evaluate the remaining terms of Eq. 7. When performing our QM calculations, we used the mean solvent potential, averaged over 200MD steps (i.e. m = 200). All QM calculations were performed using the G03 software package[214], the 6-31+G* 5D basis set and the MPW1PW91 hybrid functional[215]. The Merz-Kollman scheme[216] with default atomic radii was used to determine charges on atoms to be used later in the MD simulations. Finally, the combined QM/MM calculations were performed using a specially adapted version of MOLARIS (as outlined in Ref.62), with the Gaussian03-MOLARIS communication being facilitated by Perl scripts based on the Gaussian Output Tools package[217].

III. Specific Examples

III.1. Calculating the Sidechain pK_a of Asp3 in the Bovine Pancreatic Trypsin Inhibitor (BPTI)

Our initial test case for the validity of our QM/MM-FEP model was to evaluate the pK_a of the Asp3 residue in the bovine pancreatic trypsin inhibitor (BPTI). This is a small protein, comprised of only 58 residues, and the residue of interest lies fairly close to the surface of the protein. The atomic positions were taken from the X-ray study of BPTI reported in 1975 by Diesenhofer and Steigenmann at 1.5Å resolution[218], and the pK_a value of this residue has been determined by nuclear magnetic resonance studies[219] to be ~4.0 (compared to a pK_a^w of 3.9). The position of this residue on the surface of BPTI is shown in Fig. 4. In this figure, the protein is shown in violet, Asp3 is coloured by atom type, and all electroneutral groups within 5Å of this residue (which were explicitly defined in the QM/MM calculation) are shown in magenta.

Before evaluating the energy change involved in ionizing an acid in a protein, it is important to consider the cost of the ionisation of the corresponding acid in water. The solvation free energy for this process has been discussed in detail elsewhere[134,152], where it was demonstrated the solvation free energy of the ionized acid in water is roughly −80 kcal/mol (the experimental value for the acetate ion / aspartate is −80.7 kcal/mol[220]). The solvation free energy of the unionised acid in water is roughly −10 kcal/mol. Since the observed pK_a values are quite similar in water and proteins (pK_a^p and pK_a^w), we expect $\Delta\Delta G_{solv}$ to be approximately −70 kcal/mol both in water and in the protein.

The most challenging part in calculations of the free energy of solvation is the correct evaluation of the terms contributing to Eq. 6. In order to accurately obtain the values of $\langle\ \rangle_{E(Q)}$ and $\langle\ \rangle_{E(Q^0_{COSMO})}$, it was necessary to perform the sampling over production runs of 250 and 500ps in solution and protein respectively. Fig. 5 shows the sum of the LRA terms of Eq. 6 over the course of the production run for all systems studied. From this figure it can be seen that even in the protein, Eq. 6 converges within the first ~75 ps of the simulation, and, in all cases, convergence is to within 1 kcal/mol. The energy breakdown for the overall free energies of solvation of the acetate ion in water and the Asp3 sidechain of BPTI is shown in Table 2. Here, our values for both the acetate ion in water as well as for Asp3 in BPTI are in good agreement with what would be expected from experimental values. Thus, we can substitute the solvation free energies shown in Table 2 for the relevant values in Eq. 13 to obtain an intrinsic pK_a of 5.6 (based on a value of 3.9 for pK_a^w). We have also calculated $\Delta pK_a^{\text{ch arg es}}$ separately, for which we obtain a value of −0.8. Substituting these values back in Eq. 14 gives us an apparent pK_a of 4.8 for this residue, which is within ~1 pK_a unit of the experimentally expected value for pK_a^p (i.e. an error of 1 kcal/mol).

III.2. Calculating the Sidechain pK_a of Lys102 in T4-Lysozyme Mutant

Having obtained promising results for a surface residue on BPTI, we have extended our validation study to evaluate the pK_a of Lys102 in the M102K T4-lysozyme mutant[221]. The M102K crystal structures (obtained at pH 6.8 with1.9Å resolution) used is very similar to the wild type T4 lysozyme, with the only significant difference being the increased mobility of Lys102 and the Glu108 – Gly113 α-helix relative to the WT enzyme. The pK_a of Lys102 in this mutant form of T4-Lysozyme was measured by NMR and differential titrations[221] to have a large downward shift of four pK_a units (i.e. a pK_a^p of 6.5 relative to a pK_a^w of 10.5 for lysine), corresponding to a significant destabilisation of between 2 to 9 kcal/mol over a pH range of 10 to 3 respectively[185].

Fig. 6 shows the position of Lys102 in a hydrophobic pocket of the mutant T4-lysozyme[221]. Evaluating the pK^a of a buried residue presents a particular challenge, as unlike with surface residues, the protein can be expected to undergo a significant conformational response to a change in the protonation state. Such a major conformational reorganisation was indeed observed in a previous QM/MM-FEP study of the M102K mutant, which, while obtaining a pK_a shift in the right direction, overestimated this shift by up to 11.6 pK_a units[185] (i.e. ~16 kcal/mol). This problem occurred despite running simulations of >10ns for each λ frame, and is most likely due to the fact that the approach used inherently constrains the system such that it cannot properly respond to the protonation-induced conformational change.

Fig. 7 shows the sum of the LRA terms of Eq. 6 over the course of the production run for all systems studied (methylamine in water and the Lys102 sidechain, both neutral and charged). From this figure, it can be seen that this sum again converges quite rapidly to within 1 kcal/mol. The energy breakdown for the overall free energies of solvation of methylamine in water and the Lys102 sidechain of lysozyme are shown in Table 3. Once again, we have substituted

the solvation free energies shown in Table 3 for the relevant values in Eq. 6 to obtain an intrinsic pK_a of 3.9 (based on a value of 10.4 for pK_a^w), which, combined with our very small calculated value of $\Delta pK_a^{\text{ch arg es}}$ (0.10 in this case) gives us an apparent pK_a of 4.0. This is within 2.4 pK_a^p units of the experimental value, with the shift in the right direction (experimentally, a downward shift of 4 pK_a units is obtained[221] and here we obtain a downward shift of 6.4 pK_a units, i.e. an error of ~3kcal/mol). This is clearly a great improvement on the previously calculated value for this residue[185], and is unaffected by the fact that we are evaluating the sidechain pK_a for a residue that is buried deep in a hydrophobic surface rather than a surface residue for which the calculation is expected to be far more straightforward (as there is a much smaller likelihood of significant conformational responses to the change in protonation state). Our obtained error may be a reflection of several factors. Firstly, the classical energetics in solution could be overestimated due to the fact that the solute-solvent van der Waals parameters were not refined for the COSMO charges. Secondly, the energetics in the protein could be underestimated since we have not used a polarisable force field (see the references cited earlier for the induced dipole contributions). In fact, much more reliable results for pK_a calculations have been obtained from our earlier classical calculations[179], but this is not the point of the present work, where instead we focus on exploring the performance of proper QM/MM-FEP calculations. Also, note that in the related cases of enzyme catalysis, we will be dealing with energy differences of ~7 kcal/mol. Thus, an error of 3 kcal/mol falls within an acceptable range in order to determine the main contributions to the catalytic energy of enzymes.

III.3. Additional Examples

While our main focus here is the use of QM/MM-FEP for pKa calculations in proteins, it is important to briefly consider some other relevant examples.

III.3a. Solvation Calculations—The appreciation of the importance of properly sampling QM(ai)/MM surfaces has led to several advances in this direction[14,15,19,56–62,222]. A major focus of several of these studies[19,58–61,222] has been based on different variations on the idea[15,56] of using a classical potential as a reference for QM/MM calculations. Most notably, a recent work introduced an approach for accelerated QM(ai)/MM-FEP calculations[62] by means of an averaging potential in which the average effect of the fluctuating solvent charges is accounted for by using equivalent charge distributions, which are updated every *m* steps. This approach was then rigorously examined by evaluating the solvation of a water molecule and formate ion in water, as well as the dipole moment of water in a water solution. Here, several models for the representation of the solvent were tested in terms of accuracy and efficiency, and particular attention was paid to the convergence of the calculated solvation free energies and the corresponding solute polarization (an example of this is demonstrated in Fig. 8[62]). The most effective model was found to be one in which the system is divided into an inner region with *N* explicit solvent atoms, and an external region with two effective charges (as illustrated in Fig. 2a). However, different models were considered in terms of both the division of the solvent system and the update frequency, and, remarkably, it was found that different averaging potentials eventually converge to the same value, though some provide more optimal ways to obtain the final QM(ai)/MM converged results than others (a complete discussion and proof of this point can be found in Ref.62). This approach was demonstrated to allow for a computational time saving of up to 1000x over calculations that evaluate the QM (ai)/MM energy at every time step, while obtaining properly converging results, thus making an excellent example of a convergence study.

A practical application of this approach was demonstrated in a study of the stability of different anionic tautomers of uracil[202]. It is believed that the anionic states of nucleic acid bases play a role in the radiation damage processes of DNA[223], making them the subject of intensive experimental and theoretical studies[224–233]. Recent studies have suggested that excess

electron attachment to the nucleic acid bases can stabilize some rare tautomers such as imine-enamine tautomers and other tautomers in which a proton is being transferred from nitrogen to carbon sites[234–236]. However, early computational studies[234,235] on these compounds did not focus on accurately predicting the stability of the important anionic tautomers of nucleic acid bases in solution (that is, the relative stability of the most stable tautomers was only estimated at the DFT level, with the solvent effects being simulated by means of continuum models). Recently, the use of accelerated QM(ai)/MM-FEP simulations using an averaging potential as outlined above demonstrated that three of the recently identified anionic tautomers are 6.5– 3.6 kcal mol^{-1} more stable than the anion of the canonical tautomer while obtaining convergent results[202]. Thus, this approach is easily applicable to the study of biologically relevant reactions in solution.

III.3b. Redox Calculations—Calculations of redox and electron transport properties by classical force fields and by simplified solvent models have been quite effective at providing detailed insights into redox potentials and reorganization energies[165,166,169,172,237–242]. The treatment of the charge of the redox centers[243] and even the coupling between redox centers[244,245] has been effectively achieved by semi-empirical approaches. However, the current challenge to computational chemists is the correct evaluation of redox potentials and reorganization energies by means of QM(ai)/MM approaches, and some advances along this line will be mentioned below.

A recent computational study[57] has reported detailed calculations of the reduction potentials of the blue copper proteins plastocyanin and rusticyanin by means of a QM/MM all-atom FDFT method (which was introduced in Section I), in which the reaction centre and its closest residues and water molecules are treated by an *ab initio* approach, whereas the protein residues further away are represented with a classical forcefield. This study manages to reproduce the difference between the reduction potentials of the two blue copper proteins in a reasonable way[57] (obtaining, however, more quantitative results by the classical approach), and demonstrates that the protein permanent dipoles tune *down* the reduction potential for plastocyanin in comparison to the active site in regular water solvent, whereas in the case of rusticyanin, the reduction potential is instead tuned *up*. Also, the electrostatic environment, which is the major effect determining the reduction potential, is a property of the entire protein and solvent system, and thus cannot be ascribed to any particular single interaction[57]. It should be noted that this work was done before the implementation of the present average potential approach, and thus has not used the powerful cycle of Fig. 1. However, the FDFT approach appears to provide a consistent and effective way for reproducing the configurational ensembles needed for consistent *ab initio* free energy calculations.

Finally, the evaluation of reorganization energies by QM(ai)/MM approaches was also recently reported by several workers[246,247]. These studies were based on the formulation introduced by Warshel[248], which has been effectively used in classical studies[238,248]. More progress along this line is clearly expected.

III.3c. The Surface for Phosphate Ester Hydrolysis—Phosphate ester and anhydride hydrolysis is ubiquitous in biology, and is involved in, amongst other things, signal transduction, energy production, the regulation of protein function, as well as many metabolic and signaling pathways[249–251]. Thus, over the past few decades, significant effort has been invested into attempting to understand phosphate hydrolysis[18,252–265]. However, despite the intensive effort in the field, the precise nature of both the solution and enzyme-catalyzed reactions remain highly controversial, and attempts to imply that such controversy does not exist are highly unjustified[266]. Additionally, recent theoretical studies have demonstrated that key physical organic chemistry approaches to elucidate reaction mechanisms are in actual fact ambiguous and cannot be used to reach any unique mechanistic conclusions[18,262,264,265].

The transition states for phosphate hydrolysis reactions have traditionally been classified as either associative or dissociative[267,268], according to the distance between the reacting phosphate and the leaving group. The controversy with regards to the nature of the transition state for phosphate hydrolysis arises from the fact that the reaction can in principle proceed through multiple pathways (for detailed discussion of the potential mechanisms see Refs.18 and [262–265]), and the energies of the transition states for these pathways are relatively high making it very difficult to characterize such transition states experimentally *via* key intermediates. Thus, the only decisive way to determine the preferred reaction mechanism is by means of theoretical studies of the full reaction surface, an approach that allows for the fact that there are multiple mechanistic possibilities for phosphate hydrolysis, and allows for their direct comparison. Several such studies have already been performed, addressing some of the key mechanistic controversies in the field[18,262,264,265], and an example of such a surface is shown in Fig. 9.

To date, these works have been predominantly studies of the surface for the solution reaction, which were generated by use of implicit solvation models. However, when studying enzyme-catalyzed reactions in particular, it is necessary to use explicit solvent molecules rather than an implicit solvation model in order to be able to accurately model the interaction of the substrate with the surrounding system, and, ultimately, the key controversies in this field are likely to only be resolved by QM/MM and related calculations. In fact, it has become quite popular to study enzymatic reaction mechanisms by means of QM/MM energy minimization, such as some recent studies on DNA Polymerase β (Pol β) catalysis and fidelity[50,269–271]. However, such un-calibrated and un-validated studies are problematic on several counts, even when they manage to serendipitously reproduce the observed reaction barrier. Firstly, as was also discussed in the introduction, configurational sampling is very important as one usually gets different barriers with different energy minimized starting points. However, if QM/MM (even QM(ai)/MM) is performed with proper sampling, the obtained barrier is independent of the starting structure, thus circumventing this problem[17,62], as was demonstrated in Refs. 45 and [215]. Additionally, QM/MM studies in proteins need to be validated by first taking into account the reference reaction in water, as was done in a recent careful study of a part of the Pol β mechanism by Xiang *et al*[78]. Finally, even in the case of enzymatic reactions, it is important to take into account the full reaction surface (after of course carefully examining the reference reaction in solution) in order to be able to make conclusive decisions with regards to the preferred pathway.

One such step in this direction was taken by a recent computational study that generated the free energy surface for the GTPase reaction of the RasGAP system by means of *ab initio* QM/MM free energy calculations[18], demonstrating that whilst the overall surface is quite flat, the lowest transition state is associative. Ultimately, however, the accuracy of QM/MM approaches is not measured by the basis set, level of theory, simulation length, software used, and other such arbitrary factors, but rather in the ability of the approach to accurately reproduce relevant biochemical observables, such as pK_as. To date there is not a single work that has successfully reproduced such information by means of QM/MM energy minimization, an issue that should not be a problem with a free energy approach that uses correct sampling, as was demonstrated here. At present, further detailed QM(ai)/MM-FEP studies of the mechanism of phosphate hydrolysis both in solution and in related enzymatic systems by means of our accelerated QM(ai)/MM approach are currently underway in our research group. A preliminary example of such studies is shown in Fig. 10, which shows a comparison of 1D-reaction profiles obtained by COSMO and QM/MM-FEP for hydroxide attack on a 4-nitro substituted methyl phenyl phosphate diester.

IV. General Comments

Evaluating electrostatic energies in proteins presents a major bottleneck in the process of quantitative structure-function correlation: that is, even though we are obtaining increasingly improved high-resolution structures of complex enzymatic systems[272] (thus allowing us greater insights into the structure-functional relationships[273,274]), accurately calculating of the pK_as of ionisable groups in proteins still remains a major computational challenge. However, this issue is of key importance, not only due to the central role of ionisable groups in maintaining the structure and function of biomolecules, but also due to the fact that the availability of reliable experimental information makes pK_a calculations an important measure of the accuracy of electrostatic calculations in general and QM/MM calculations in particular. Although the crucial need for sampling QM/MM calculations has been highlighted by empirical valence bond studies[42,43] as early as 1980, the issue of the accurate treatment of electrostatics for QM/MM methods is one that is only recently being addressed by the wider scientific community[25,62,185,186,189,275–277]. The importance of proper sampling is particularly challenging when one uses *ab initio* QM representations in QM/MM calculations, and this issue is the main point of the present work.

Solution pK_as have been evaluated using experimental gas phase energies with calculated solvation energies[278], as well as by using quantum mechanical calculations of gas phase energies with a macroscopic estimate of the solvation energy[279,280] and also by a combination of gas-phase *ab initio* charges and FEP calculations[281]. However, this is almost trivial in comparison to the challenges involved in evaluating pK_as in proteins. For instance, in solution, one can obtain an almost perfect agreement between calculated and experimental pK_a values merely by calibrating the empirical van der Waals[278] or Born radii, whereas in a protein, the pK_a is different in a different region so a pre-defined radius cannot consistently reproduce the correct value for the entire protein. Correctly evaluating free energy changes in proteins (and also in solution) by classical MM or by QM/MM approaches often requires very extensive averaging over the configurational space of the protein. Thus, merely performing simple minimization as is done in gas-phase QM calculations is not sufficient in order to effectively evaluate the activation energies of chemical reactions in proteins[49]. However, evaluating the free energies of QM(ai)/MM surfaces is extremely challenging due to the need for the extensive evaluation of the QM(ai) energies. Therefore, even though several innovative strategies have already been suggested for accelerating the QM(ai)/MM sampling[14,15,56,64,65,198], there is a clear need for more “mainstream” approaches than can help obtain converging QM(ai)/MM free energies (particularly for examining free energy changes in proteins) within reasonable computational cost.

The approach we present here is an extension of our previous work in solution[62,217] and has been extensively validated for small model systems[62]. Our protocol for obtaining free energies is effectively a two-step approach: first, we perform classical MD simulations in order to obtain the classical solvation free energy by means of the free energy perturbation adiabatic charging (FEP-AC) procedure. Here, it is assumed that the solvated molecules have the charge distributions obtained by standard *ab initio* continuum approaches (or, in the case of the protein, by a few steps of QM/MM). We subsequently refine the free energy of solvation by taking into account the real, average protein/solvent charge distribution over the course of a QM/MM simulation, which reflects the polarisation that is caused by the explicit water molecules used in our solvation model. This was obtained by use of our accelerated QM/MM simulation (where the QM energy of the solute is evaluated in the mean solvent potential, with an averaging every 200 MD steps). We have used this method to evaluate the sidechain pK_as of Asp3 and Lys102 in the bovine pancreatic trypsin inhibitor (BPTI) and the M102K T4-lysozyme mutants respectively. In both cases, we obtain values that lie within ~2.5 pK_a units (i.e. 3 kcal/mol) of the experimental value.

Our result for Lys102 in the T4 lysozyme is of particular interest. Here, we are examining a charged residue that is positioned well into a hydrophobic pocket in the protein[221]. A previous QM/MM-FEP study[185] obtained a very large error in the pK_a shift (up to 11.6 pK_a units, i.e. approximately 16 kcal/mol), despite running simulations of >10ns per frame. It was argued that the difficulty in obtaining an accurate result for this residue was due to the fact that whilst Lys102 is deprotonated in the X-ray structure, the protein is highly destabilized when this residue is protonated resulting in a major conformational re-organization. Even though such a significant conformational response to a change in the protonation state may be expected for most buried residues, the approach used in this study only allows for limited conformational flexibility. The relationship between pK_a and protein conformation has been discussed in detail elsewhere[282–286]. That is, it has been argued that the large differences between pK_as calculated using different crystal structures indicate the importance of conformational effects[287]. However, this overlooks the fact that the pKa simply reflects the average effect (free energy) of all relevant conformations. Therefore, the important issue is to obtain a reliable averaging over different conformations. While such an averaging will not necessarily improve the obtained values, it will provide results that are more robust. Thus, it is important to have correct averaging over both the ionised and neutral states. If this exists, a situation such as the conformational re-organization of lysozyme upon the protonation of the lysine should not significantly affect the calculated pK_a value. We have demonstrated that this is the case with the protocol here: regardless of whether we are examining a surface residue or one buried deep in the hydrophobic pocket, we do not obtain an error greater than ~3 kcal/mol. Additionally, even though the error obtained here may seem large, it is important to bear in mind that when examining enzyme catalysis, we are dealing with energy differences of up to 7 kcal/mol. As such even an error margin of 3 kcal/mol is sufficient in order to be able to determine the main contributions to the 7 kcal/mol catalytic energy, thus providing a useful tool for structure-function correlation studies. Therefore, we believe that our accelerated QM/MM is a highly efficient and powerful tool to predict the electrostatics of not only solution but also enzymatic reactions, as well as the solvation free energies of even larger systems, such as nucleic acid bases incorporated into DNA[202], and studies to demonstrate this are already underway in our group.

It should be noted at this point that there has recently been an interesting study[288] that uses a thermodynamic integration approach to calculate the pK_a of residue 66 in two *Staphylococcal nuclease* mutants (V66E and V66D). Amongst other things, this work addresses a similar issue as our earlier study[289], namely the challenge of evaluating the pK_as of groups that are located in non-polar regions in proteins, where macroscopic models would predict a much larger pK_a shift than that observed experimentally[290–292]. Our work[289] (that used a novel overcharging approach to accelerate water penetration and local unfolding) reproduced the observed pK_a of V66E with small local rearrangements and some water penetration. On the other hand, Ref.288 concluded that E66 moves spontaneously to the solvent region within 500ps of the 6ns total simulation time. This work involves some misunderstandings. First, in contrast to the claims of Ref.288, we never suggested any *major* conformational changes in our original study, but rather we proposed limited local unfolding (as shown in Fig. 5 of Ref. 289). Furthermore, the conformational changes of E66 found in Ref.288 where first identified by our overcharging approach (again see Fig. 5 of Ref.289), but not as a minimum on the free energy surface of the ionized residue. The difference may well be due to the extra stabilization of the ion pair formed between E66 and Lys63 in the simulations performed in Ref.288. In our view, the ion pair effect cannot explain the observed pK_a (a point supported by related experiments[293]). Probably, the ion pair effect reflects non-perfect boundary conditions that do not provide a high enough dielectric constant for the ion pair, and this issue will have to be explored in subsequent studies. In fact, it should be pointed out that we have invested extensive effort into reproducing these results, and we neither observe a dramatic conformational change, nor any significant effect due to the presence of the ion pair. Nevertheless, this recent work by

Ghosh and Cui[288] demonstrates that nice progress is being made towards addressing more challenging pK_a problems.

V. Perspectives

The ability to perform QM(ai)/MM calculations of solvation free energies provides an important departure from studies that involve continuum models. That is, there are key questions in biochemistry whose resolution critically depends on the nature of the reference solution reaction. An excellent example is the mechanism of phosphate hydrolysis in solution and the reaction of the ribosome[294–300] as well as other key reactions. In each of theses cases there is major controversy about the nature of the reference reaction where recent theoretical studies prove conclusively that the key physical organic chemistry approaches used to resolve mechanistic problems cannot provide any unique mechanistic conclusions[18,262,264,265]. This means that the key controversies in the field will eventually be resolved by QM/MM and related calculations. Here one can start with simplified solvation models (e.g. continuum models such as PCM, COSMO or the LD model) but this led to controversies and confusion, including the recent suggestion to use a mixed solvation model where the calculations are performed in a continuum with a certain number of explicit water molecules added to the system[301,302]. However, this approach is hugely problematic on several counts. For example, the presence of explicit water molecules will make accurate transition state determination without mapping the full surface virtually impossible, as the non-bonding interactions brought about by the loosely bound species introduced into the system will create a large number of soft vibrations that could seriously complicate the accurate identification of the correct transition state[262,263]. Of great importance when examining a solvated system is to have correct boundary conditions for the interaction between the solute and the bulk, and it is questionable whether solvation models which mix continuum and explicit water accurately reproduce these boundary conditions (or, indeed take into account the entropic contribution of the explicit water molecules to the overall free energy barrier). The issue of proper polarisation boundary conditions between the explicit and continuum regions has been explored and emphasized in many of our works[133,213,278], and has also received recognition recently by workers other than us[303]. Basically, the problem is that the water molecules in the first solvation shells will be over-polarised if they are not subjected to polarisation constraints that properly represent the effect of the rest of the system in the true infinite system. This situation becomes in turn more serious when one only uses a few water molecules immersed in a continuum model. Here, the explicit water molecules are not likely to have the correct orientation of the corresponding molecules in an explicit infinite system. This issue can be easily tested by simply taking four water molecules in a continuum and checking whether the minimized structure has any similarity to the corresponding simulated structure of water in a large simulation system. In a similar problem, three water molecules near an ion will always be over-polarised relative to the polarisation of the nearest neighbors in a large water sphere.

Perhaps a more serious problem is the fact that the inclusion of several explicit water molecules in a continuum model combined with an energy minimization treatment would require the evaluation of the entropic effect associated with the explicit solvent molecules (this effect is included implicitly in the continuum model) , an issue which has effectively been ignored by previous studies of phosphate hydrolysis that utilize such an approach[301,302,304]. Therefore, unless one is specifically interested in the chemistry (e.g. proton transfer) associated with the presence of a few explicit water molecules, having a mixed model is problematic and it seems to us that one has to either use a well calibrated simplified model (such as Langevin Dipoles or Poisson-Boltzmann) or a QM/MM model with a proper convergent free energy treatment[62]. These requirements are covered by our approaches. Additionally, including a polarisable force field in the description of the solvent is also rather trivial and has been implemented in many of our treatments[112,134,152,153,305]. What is still missing in the

current model is a proper treatment of charge transfer to the solvent. This feature can be incorporated in the CDFT approach[73], but more studies are needed to establish the viability of this approach.

As has been seen here, hybrid quantum mechanical / molecular mechanical simulations are becoming an increasingly popular tool for understanding biochemical and biophysical systems on a molecular level, and significant progress has been made in this direction over the past few years. Of particular interest are the FDFT/CDFT approaches which allow for the entire protein to be represented quantum mechanically, as well as progress in reliable *ab initio* QM/MM free energy perturbation calculations with proper sampling. At present, the limiting factor in both cases is computational cost. However, as computers become increasingly more powerful, this is an issue that will become less and less of a problem, and this in turn will ultimately allow us to evaluate the potential of mean force (PMF) for enzymatic reactions by QM(ai)/MM approaches, which is currently extremely challenging[49,62] due to the requirement of very extensive sampling which results in the computationally expensive repeated evaluation of the QM/MM energies. Additionally, earlier work has cast an interesting light on the relationship between solvent fluctuations and the convergence of QM(ai)/MM calculations[62], which can be exploited in QM(ai)/MM free energy calculations that do not keep the solvent fixed but rather allow the solvent to fluctuate.

The idea of using a reference potential in sampling complex[88,306] or very expensive[56,72] surfaces is becoming a powerful tool in QM/MM studies and is likely to be explained in mainstream approaches either in its current form or in the closely related metadynamics method. Here, the issue of choosing the best reference potential can be complicated. In the case of solvation problems, the best reference is obviously the given solute with a reasonable charge set. In the case of chemical reactions, we believe that the current best reference potential is the EVB potential and the best reaction coordinate is the EVB energy gap. However, looking for alternatives may result in more effective treatments.

Essentially, we are dealing here with a new and rapidly emerging field with many users presenting different approaches. Therefore, judgments with regards to the accuracy of the work may be skewed by statements about ever increasing simulation lengths, the level of theory used, and comparisons to other workers who obtain similar results. Here, there is clearly a necessity for collective education, where it is essential to realize that *only* systematic validation studies on seemingly auxiliary results like pK_as can be used to establish the accuracy of a method, and, additionally, only studies, which examine complete energy surfaces, can be used to establish conclusive mechanistic results. Thus, amidst all the potential confusion with regards to whom to believe (which risks giving highly qualified workers in the field a Cassandra complex), those who can correctly reproduce pK_as and reference reactions in solution should be more believable than those who reproduce high level gas-phase SCF results (though of course this is not the case with gas-phase *ab initio* calculations where the results have become widely accepted, even by experimentalists). Ultimately, our hope is not only that QM/MM approaches become more widely used, but also that the understanding that the quality of the results does not only depend on technological aspects will slowly emerge for both the theoretical and the experimental communities.

Acknowledgements

This work was supported by NIH grant GM22492, NIH Grant GM40283 and NSF Grant MCB-0342276. All computational work was supported by the University of Southern California High Performance Computing and Communication Centre (HPCC). Maciej Haranczyk is a 2008 Seaborg Fellow at Lawrence Berkeley National Laboratory. This research was supported in part (to M. H.) by the U. S. Department of Energy under contract DE-AC02-05CH11231. We would also like to thank Spyridon Vicatos for insightful discussion, and Robert Rucker for his assistance in the preparation of the manuscript.

References

1. Cui Q, Elstner M, Kaxiras E, Frauenheim T, Karplus M. J. Phys. Chem. B 2001;105:569–585.
2. Monard G, Merz KM. Acc. Chem. Res 1999;32:904–911.
3. Shurki A, Warshel A. Adv. Protein. Chem 2003;66:249–312. [PubMed: 14631821]
4. Warshel A, Levitt M. J. Mol. Biol 1976;103:227–249. [PubMed: 985660]
5. Gao J. Acc. Chem. Res 1996;29:298–305.
6. Bakowies D, Thiel W. J. Phys. Chem 1996;100:10580–10594.
7. Field MJ, Bash PA, Karplus M. J. Comp. Chem 1990;11:700–733.
8. Friesner RA, Beachy MD. Curr. Op. Struct. Biol 1998;8:257–262.
9. Garcia-Viloca M, Gonzalez-Lafont A, Lluch JM. J. Am. Chem. Soc 2001;123:709–721. [PubMed: 11456585]
10. Marti SAJ, Moliner V, Sille E, Tunon I, Bertran J. Theor. Chem. Acc 2001;3:207–212.
11. Field M. J. Comp. Chem 2002;23:48–58. [PubMed: 11913389]
12. Lyne PD, Mulholland AJ, Richards WG. J. Am. Chem. Soc 1995;117:11345–11350.
13. Warshel A, Sussman F, Hwang J-K. J. Mol. Biol 1988;201:139–159. [PubMed: 3047396]
14. Zhang Y, Liu H, Yang W. J. Chem. Phys 2000;112:3483–3492.
15. Muller RP, Warshel A. J. Phys. Chem. B 1995;106:13333–13343.
16. Aqvist J, Warshel A. Chem. Rev 1993;93:2523–2544.
17. Rosta E, Klähn M, Warshel A. J. Phys. Chem. B 2006;110:2934–2941. [PubMed: 16471904]
18. Klahn M, Rosta E, Warshel A. J. Am. Chem. Soc 2006;128:15310–15323. [PubMed: 17117884]
19. Rod TR, Ryde U. Phys. Rev. Lett 2005;94:138302. [PubMed: 15904045]
20. Wladkowski BD, Krauss M, Stevens WJ. J. Phys. Chem 2001;99:6273–6276.
21. Crespo A, Scherlis A, Marti MA, Ordejon P, Roitberg AE, Estrin DA. J. Phys. Chem. B 2003;107:13728–13736.
22. Shurki A, Warshel A. Proteins 2004;55:1–10. [PubMed: 14997535]
23. Futugasi N, Hata M, Hoshino T, Tsuda M. Biophys J 1999;77:3287–3292. [PubMed: 10585950]
24. Glennon TM, Villa J, Warshel A. Biochemistry 2000;39:9641–9651. [PubMed: 10933780]
25. Warshel A. Annu. Rev. Biophys. Biomol. Struct 2003;32:425–443. [PubMed: 12574064]
26. Florian J, Goodman MF, Warshel A. Biopolymers 2003;68:286–299. [PubMed: 12601790]
27. Cavalli A, Carloni P. J. Am. Chem. Soc 2002;124:3763–3768. [PubMed: 11929266]
28. Langen R, Schweins T, Warshel A. Biochemistry 1992;31:8691–8696. [PubMed: 1390653]
29. Schweins T, Langen R, Warshel A. Nature Struct. Biol 1994;1:476–484. [PubMed: 7664067]
30. Chung HH, Benson DR, Schultz PG. Science 1993;5:806–809. [PubMed: 8430333]
31. Schweins T, Geyer M, Scheffzek K, Warshel A, Kalbitzer HR, Wittinghofer A. Nature Struct. Biol 1995:36–44. [PubMed: 7719852]
32. Friesner R, Dunietz B. Acc. Chem. Res 2001;34:351–358. [PubMed: 11352713]
33. Bloomberg MRA, Siegbahn PEM. J. Phys. Chem 2001;105:9376–9386.
34. Pelmenshikov V, Siegbahn PEM. Inorg. Chem 2005;44:3311–3320. [PubMed: 15847441]
35. Blomberg LM, Blomberg MRA, Siegbahn PEM. J. Inorg. Biochem 2005;99:949–958. [PubMed: 15811512]
36. Borowski T, Bassan A, Richards NG, Siegbahn PEM. J. Chem. Theory. Comput 2005;1:686–693.
37. Kim J, Tsai P-C, Chen S-L, Himo F, Almo SC, Raushel FM. Biochemistry 2008;47:9497–9504. [PubMed: 18702530]
38. Georgieva P, Himo F. Chem. Phys. Lett 2008;463:214–218.
39. Peeters A, Swerts B, Alsenoy CV. J. Phys. Chem. B 2003;107:4871–4878.
40. Liu XH, Sun M, Yue JJ, Yin YX. Journal of Molecular Structure: THEOCHEM 2003;620:227–230.
41. Sharma PK, Chu ZT, Olsson MH, Warshel A. Proc. Natl. Ada. Sci. U. S. A 2007;104:9661–9666.
42. Warshel A, Weiss RM. J. Am. Chem. Soc 1980;102:6218–6226.
43. Warshel A. Proc. Natl. Acad. Sci. U.S.A 1984;81:444–448. [PubMed: 6582500]

44. Stanton RV, Perakyla M, Bakowies D, Kollman PA. J. Am. Chem. Soc 1998;120:3448–3457.
45. Kollman PA, Kuhn B, Donini O, Perakyla M, Stanton R, Bakowies D. Acc. Chem. Res 2001;34:72–79. [PubMed: 11170358]
46. Chandrasekhar J, Jorgensen WL. J. Am. Chem. Soc 1984;106:3049–3059.
47. Hwang JK, King G, Creighton S, Warshel A. J. Am. Chem. Soc 1988;110:5297–5311.
48. Warshel, A. Computer modeling of chemical reactions in enzymes and solutions. New York: John Wiley and Sons; 1991.
49. Klähn M, Braund-Sand S, Rosta E, Warshel A. J. Phys. Chem. B 2005;109:15645–15650. [PubMed: 16852982]
50. Alberts IL, Wang YA, Schlick T. J. Am. Chem. Soc 2007;129:11100–11110. [PubMed: 17696533]
51. Grigorenko BL, Nemukhin AV, Topol IA, Cachau RE, Burt SK. Proteins: Struct. Funct. Bioinf 2005;60:495–503.
52. Yang Z, Zhao Y-P. Materials Science and Engineering A 2006;423:84–91.
53. Mulholland AJ, Lyne PD, Karplus M. J. Am. Chem. Soc 2000;122:534–535.
54. Dittrich M, Hayashi H, Schulten K. Biophys J 2003;85:2253–2266. [PubMed: 14507690]
55. Schöneboom JC, Cohen S, Lin H, Shaik S, Thiel W. J. Am. Chem. Soc 2004;126:4017–4034. [PubMed: 15038756]
56. Strajbl M, Hong G, Warshel A. J. Phys. Chem. B 2002;106:13333–13343.
57. Olsson MH, Hong G, Warshel A. J. Am. Chem. Soc 2003;125:5025–5039. [PubMed: 12708852]
58. Liu WW, Doren DJ. J. Phys. Chem. B 2003;107:9505–9513.
59. Iftimie RS, Schofield J. J. Chem. Phys 2003;119:11285–11297.
60. Crespo AM, Estrin DA, et al. J. Am. Chem. Soc 2005;127:6940–6941. [PubMed: 15884923]
61. Pradipta B. J. Chem. Phys 2005;122:91102.
62. Rosta E, Haranczyk M, Chu ZT, Warshel A. J. Phys. Chem. B 2008;112:5680–5692. [PubMed: 18412414]
63. Hu H, Lu ZY, Yang WT. J. Chem. Theor. Comput 2007;3:390–406.
64. Ishida T, Kato S. J. Am. Chem. Soc 2003;125:12035–12048. [PubMed: 14505425]
65. Ruiz-Pernia JJ, Silla E, Tunon I, Marsi S, Molliner V. J. Phys. Chem. B 2004;108:8427–8433.
66. Kaukonen M, Soderhjelm P, Heimdal J, Ryde U. J. Chem. Theory Comput 2008;4:985–1001.
67. Strajbl M, Florian J, Warshel A. J. Am. Chem. Soc 2000;122:5354–5366.
68. Bentzien J, Muller RP, Florian J, Warshel A. J. Phys. Chem. B 1998;102:2293–2301.
69. Truhlar DG. J. Phys. Chem. A 2002;106:5048–5050.
70. Yang W. Phys. Rev. Lett 1991;66:1438–1441. [PubMed: 10043209]
71. Lee TS, Yang W. Int. J. Quant. Chem 1998;69:397–404.
72. Wesolowski T, Muller RP, Warshel A. J. Phys. Chem 1996;100:15444–15449.
73. Hong G, Strajbl M, Wesolowski T, Warshel A. J. Comp. Chem 2000;21:1554–1561.
74. Cortona P. Phys. Rev. B 1991;44:8454–8458.
75. Wesolowski T, Warshel A. J. Phys. Chem 1993;97:8050–8053.
76. Wesolowski T. Phys. Rev. Lett 2002;88:209701-1. [PubMed: 12005616]
77. Kluner T, Govind N, Wang YA, Carter E. Phys. Rev. Lett 2001;86:5954–5957. [PubMed: 11415402]
78. Xiang Y, Warshel A. J. Phys. Chem. B 2008;112:1007–1015. [PubMed: 18166038]
79. Car R, Parinello M. Phys. Rev. Lett 1985;55:2471–2474. [PubMed: 10032153]
80. Tuckerman M, Marx D, Klein ML, Parinello M. Science 1997;275:817–820. [PubMed: 9012345]
81. Eichinger M, Tavan P, Hutter J, Parrinello M. J. Chem. Phys 1999;110:10452–10467.
82. Woo TK, Margl PM, Blochl PE, Ziegler T. J. Phys. Chem. B 1997;101:7877–7880.
83. Bucher, D.; Rothlisberger, U.; Guidoni, L.; Carloni, P. 228th ACS National Meeting. Philadelphia, PA: United States; 2004. p. PHYS-309
84. Yang S-Y, Ziegler T. Organometallics 2006;25:887–900.
85. Moret M-E, Tapavicza E, Guidoni L, Rohrig UF, Sulpizi M, Tavernelli I, Rothlisberger U. 2005;59:493–498.

86. Laio A, Parrinello M. Proc. Natl. Ada. Sci. U. S. A 2002;99:12562–12566.
87. Muller, RP.; Warshel, A. Ab initio calculations of free energy barriers for chemical reactions in solution: Proton transfer in [FHF]. Singapore: World Scientifc Press; 1996.
88. Fan ZZ, Hwang JK, Warshel A. Theor. Chem. Acc 1999;103:77–80.
89. Luzhkov V, Warshel A. J. Comp. Chem 1992;13:199–213.
90. Huber T, Torda AE, van Gunsteren F. J. Comput.-Aided. Mol. Design 1994;8:695–708.
91. Ensing B, De Vivo M, Liu Z, Moore P, Klein ML. Acc. Chem. Res 2006;39:73–81. [PubMed: 16489726]
92. Gervasio FL, Laio A, Parrinello M. J. Am. Chem. Soc 2005;127:2600–2607. [PubMed: 15725015]
93. Donadio D, Bernasconi M. Phys. Rev. B 2005;71:073307.
94. Asciutto E, Sagui C. J. Phys. Chem. A 2005;109:7682–7687. [PubMed: 16834142]
95. Martonak R, Laio A, Bernasconi M, et al. Zeitschrift fur Kristallographie 2005;220:489–498.
96. Urakawa A, Iannuzzi M, Hutter J, Baiker A. Chemistry 2007;13:6828–6840. [PubMed: 17566132]
97. Michel C, Laio A, Mohamed F, Krack M, Parrinello M, Milet A. Organometallics 2007;26:1241–1249.
98. Chandler, D. Finding transition pathways: Throwing ropes over rough mountain passes, in the dark". Singapore: World Scientific; 1998.
99. Dellago C, Boluis PG, Geissler PL. Adv. in Chem. Phys 2002;123:1–78.
100. Bolhuis PG, Chandler D, Dellago C, Geissler P. Ann. Rev. Phys. Chem 2002;59:291–318. [PubMed: 11972010]
101. Warshel, A.; Florian, J. Empirical valence bond and related approaches. John Wiley and Sons Ltd.; 2004.
102. Jensen F, Norrby PO. Theor. Chem. Acc 2003;109:1–7.
103. Florian J. J. Phys. Chem. A 2002;106:5046–5047.
104. Braun-Sand, S.; Warshel, A. Electrostatics of proteins: principles, models and applications. Vol. Vol.. Wiley-VCH: Weinheim; 2005.
105. Higashi M, Truhlar DG. J. Chem. Theory Comput 2008;4:790–803.
106. Albu TV, Corchado JC, Truhlar DG. J. Phys. Chem. A 2001;105:8465–8487.
107. Kim Y, Corchado JC, Villa J, Xing J, Truhlar DG. J. Chem. Phys 2000:2718–2735.
108. Hong G, Rosta E, Warshel A. J. Phys. Chem. B 2006;110:19570–19574. [PubMed: 17004821]
109. Lappe J, Cave RJ, Newton MD, Rostov IV. J. Phys. Chem. B 2005;109:6610–6619. [PubMed: 16851742]
110. Kim HJ, Hynes JT. J. Am. Chem. Soc 1992;114:10508–10528.
111. Warshel A, Parson WW. Q. Rev. Biophys 2001;34:563–679. [PubMed: 11852595]
112. Aqvist J, Warshel A. Biophys J 1989;56:171–182. [PubMed: 2473789]
113. Parsegian A. Ann. NY Acad. Sci 1975;264:161–179. [PubMed: 1062955]
114. Burykin A, Schutz CN, Villa J, Warshel A. Proteins: Struct. Funct. Genet 2002;47:265–280. [PubMed: 11948781]
115. Aqvist J, Luzhkov V. Nature 2000;404:881–884. [PubMed: 10786795]
116. de Groot BL, Frigato T, Helms V, Grubmuller H. J. Mol. Biol 2003;333:279–293. [PubMed: 14529616]
117. Jensen MO, Tajkhorshid E, Schulten K. Biophys J 2003;85:2884–2889. [PubMed: 14581193]
118. De Groot BL, Grubmuller H. Curr. Opin. Struct. Biol 2005;15:176–183. [PubMed: 15837176]
119. Chakrabati N, Roux B, Pomes R. J. Mol. Biol 2004;343:493–510. [PubMed: 15451676]
120. Miloshevsky GV, Jordan PC. Biophys J 2004;87:3690–3702. [PubMed: 15377535]
121. Ilan B, Tajkhorshid E, Schulten K, Voth GA. Proteins: Struct. Funct. Genet 2004;55:223–228. [PubMed: 15048815]
122. Burykin A, Warshel A. FEBS Lett 2004;570:41–46. [PubMed: 15251436]
123. Burykin A, Warshel A. Biophys J 2003;85:3696–3706. [PubMed: 14645061]
124. Warshel A, Weiss RM. J. Am. Chem. Soc 1981:103–446.

125. Strajbl M, Shurki A, Warshel A. PNAS 2003;100:14834–14839. [PubMed: 14657336]
126. Xiang Y, Oelschlager P, Florián J, Goodman MF, Warshel A. Biochemistry 2006;45:7036–7048. [PubMed: 16752894]
127. Warshel A, Schlosser DW. Prot. Natl. Acad. Sci. U. S. A 1981;78:5564–5568.
128. Hu Z, Ma B, Wolfson H, Nussinov R. Proteins 2000;39:331–342. [PubMed: 10813815]
129. Guerois R, Nielsen JE, Serrano L. J. Mol. Biol 2002;2002:369–387. [PubMed: 12079393]
130. Kortemme T, Baker D. Proc. Natl. Ada. Sci. U. S. A 2002;99:14116–14121.
131. Wang T, Tomic S, Gabdoulline RR, Wade RC. Biophys J 2004;87:1618–1630. [PubMed: 15345541]
132. Dong F, Zhou H-Y. Biophys J 2002;83:1341–1347. [PubMed: 12202359]
133. Warshel A, Sharma PK, Kato M, Parson W. BBA 2006;1764:1647. [PubMed: 17049320]1676
134. Warshel A, Russel ST. Q. Rev. Biophys 1984;17:283. [PubMed: 6098916]
135. Perutz MF. Science 1978;201:1187–1191. [PubMed: 694508]
136. Warshel A. Acc. Chem. Res 1981;14:284–290.
137. Sharp KA, Honig B. Annu. Rev. Biophys. Chem 1990;19
138. Nakamura H. Roles of electrostatic interactions in proteins 1996;29:1–90.
139. Davis ME, McCammon JA. Chem. Rev 1990;90:509–521.
140. Warshel A, Papazyan A. Curr. Opin. Struct. Biol 1998;8:211–217. [PubMed: 9631295]
141. Simonson T. Rep. Prog. Phys 2003;66:737–787.
142. Warshel A, Aqvist J. Annu. Rev. Biophys. Chem 1991;20:267–298.
143. Parson, WW.; Warshel, A. Calculation of electrostatic energies in proteins using microscopic, semimicroscopic and macroscopic models and free-energy perturbation approaches. Springer: Dordrecht; 2006.
144. Linderstrom-Lang K. Lab. Carlsberg 1924;15:1–29.
145. Johannin G, Kellershohn N. Biochem. Biophys. Res. Commun 1972;49:321–327. [PubMed: 4640361]
146. Hayes DM, Kollman PA. J. Am. Chem. Soc 1976;98:3335–3345. [PubMed: 1262648]
147. Tanford C, Kirkwood JG. J. Am. Chem. Soc 1957;79:5333–5339.
148. Tanford C, Roxby R. Biochemistry 1972;11:2192–2198. [PubMed: 5027621]
149. Warshel A, Russel ST, Churg AK. Proc. Natl. Acad. Sci. U.S.A 1984;81:4785–4789. [PubMed: 6589625]
150. Warshel A. Chem. Phys. Lett 1978;55:454–458.
151. Warshel A. Biochemistry 1981;20:3167–3177. [PubMed: 7248277]
152. Russel ST, Warshel A. J. Mol. Biol 1985;185:389–404. [PubMed: 2414450]
153. Lee FS, Chu ZT, Warshel A. J. Comp. Chem 1993;14:161–185.
154. Warshel A, Naray-Szabo G, Sussman F, Hwang J-K. Biochemistry 1989;28:3629–3637. [PubMed: 2665806]
155. Sham YY, Chu ZT, Warshel A. J. Phys. Chem. B 1997;101:4458–4472.
156. Gilson MK, Rashin A, Fine R, Honig B. J. Mol. Biol 1985;184:503–516. [PubMed: 4046024]
157. Gilson M, Sharp KA, Honig B. J. Comp. Chem 1987;9:327–335.
158. Warwicker J, Watson HC. J. Mol. Biol 1982;157:671–679. [PubMed: 6288964]
159. Warshel A, Papazyan A, Muegge I. J. Bioinorganic Chem 1997;2:143–152.
160. Sharp KA, Honig B. Annu. Rev. Biophys. Chem 1990;19:301–332.
161. Gilson M, Honig B. Proteins: Struct. Funct. Genet 1988;4:7–18. [PubMed: 3186692]
162. Sharp KA, Honig B. J. Phys. Chem 1990;94:7684–7692.
163. Nicholls A, Honig B. J. Comp. Chem 1991;12:435–445.
164. Honig B, Nicholls A. Science 1995;268:1144–1149. [PubMed: 7761829]
165. Gunner M, Nicholls A, Honig B. J. Phys. Chem 1996;100:4277–4291.
166. Lancaster CRD, Michel H, Honig B, Gunner M. Biophys. J 1996;70:2469–2492. [PubMed: 8744288]
167. Alexov E, Gunner MR. Biophys. J 1997;72:2075. [PubMed: 9129810]

168. Rabenstein B, Ullman GM, Knapp EW. Biochemistry 1998;37:2488–2495. [PubMed: 9485397]
169. Alexov EG, Gunner MR. Biochemistry 1999;38:8253–8270. [PubMed: 10387071]
170. Nielsen JE, Andersen KV, Honig B, Hooft RV, Klebe G, Vriend G, Wade RC. Protein Eng 1999;12:657–662. [PubMed: 10469826]
171. Ullman GM, Knapp EW. Eur. Biophys. J. Biophys. Lett 1999;38:8235–8270.
172. Gunner M, Alexov E. Biochem. Biophys. Acta 2000;1485:63–87. [PubMed: 10832090]
173. Georgescu RE, Alexov E, Gunner MR. Biophys. J 2002;83:1731–1748. [PubMed: 12324397]
174. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. J. Comp. Chem 2002;23:128–137. [PubMed: 11913378]
175. Voigt P, Knapp EW. J. Biol. Chem 2003;278:51993–52001. [PubMed: 12975370]
176. Kim J, Mao J, Gunner M. J. Mol. Biol 2005;348:1283–1298. [PubMed: 15854661]
177. Schutz CN, Warshel A. PROTEINS: Structure, Function, Genetics 2001;44:400–417.
178. Warshel A, Sussman F, King G. Biochemistry 1986;25:8368–8372. [PubMed: 2435316]
179. Lee FS, Warshel A. J. Chem. Phys 1992;97:3100–3107.
180. Lee FS, Chu ZT, Warshel A. J. Phys. Chem. B 1997;101:4457–4472.
181. Del Buono GS, Figueirido FE, Levy RM. PROTEINS: Structure, Function, Genetics 1994:85–97.
182. Brandsdal BO, Smalas AO, Aqvist J. Proteins 2006;64:740–748. [PubMed: 16752417]
183. Simonson T, Carlsson J, Case D. J. Am. Chem. Soc 2004;126:4167–4180. [PubMed: 15053606]
184. Saito M. J. Phys. Chem 1995;99:17403–17048.
185. Riccardi D, Shaefer P, Cui Q. J. Phys. Chem. B 2005;109:17715–17733. [PubMed: 16853267]
186. Li G, Cui Q. J. Phys. Chem. B 2003;107:14521–14528.
187. Pople JA. Angew. Chem.Int. Ed. Engl 1999;38:1894.
188. Bentzien, J.; Florian, J.; Glennon, T.; Warshel, A. QM/MM approaches for studying chemical reactions in proteins and solutions. Vol. Vol. 712. Washington, DC: American Chemical Society; 1998.
189. Florián J, Warshel A. J. Phys. Chem 1997;101:5538–5595.
190. Tapia O, Goscinski O. Mol. Phys 1975;29:1653–1661.
191. Cramer, CJ.; Truhlar, DG. Continuum solvation models: Classical and quantum mechanical implementations. Vol. Vol. 6. New York: VCG; 1995.
192. Tomasi J, Bonaccorsi R, Cammi R, Delvalle FJO. J. Mol. Struc 1991;80:401–424.
193. Rivali, JL.; Rinaldi, D. Computational chemistry: Review of current trends. Singapore: World Scientific Publishing; 1995.
194. Sanchez ML, Martin ME, Aguilar MA, Del Valle FJO. J. Comput. Chem 2000;21:705–715.
195. Sanchez ML, Martin ME, Galvan IF, Del Valle FJO, Aquilar MA. J. Phys. Chem. B 2002;106:4813–4817.
196. Mendoza MLS, Aguilar MA, Del Valle FJO. J. Mol. Struct 1998;426:181–190.
197. Sanchez ML, Aquilar MA, Del Valle FJO. J. Comput. Chem 1997;18:313–322.
198. Hu H, Lu ZY, Yang WT. J. Chem. Theor. Comput 2007;3:390–406.
199. Lim C, Bashford D, Karplus M. J. Phys. Chem 1991;95:5610–5620.
200. Levy RM, Belhadj M, Kitchen DB. J. Chem. Phys 1991;95:3627–3633.
201. Yang A-S, Honig B. J. Mol. Biol 1993;231:1993.
202. Haranczyk M, Gutowski M, Warshel A. Phys. Chem. Chem. Phys 2008;10:4442–4448. [PubMed: 18654684]
203. Warshel A. J. Phys. Chem 1982;86:2218–2224.
204. Lee FS, Chu ZT, Bolger MB, Warshel A. Protein Eng 1992;5:215–228. [PubMed: 1409541]
205. Barone V, Cossi M. J. Phys. Chem. A 1998;102:1995–2001.
206. Klamt A, Schüürmann GJ. J. Chem. Soc., Perkin Trans. 2 1993;5:799–805.
207. Tomasi J, Mennucci B, Cancès MT. J. Mol. Struct. (Theochem) 1991;464:464.
208. Cancès MT, Mennucci B, Tomasi J. J. Chem. Phys 1997;107:3032–3041.
209. Mennucci B, Tomasi J. J. Chem. Phys 1997;106:5151–5158.

210. Mennucci B, Cancès MT, Tomasi J. J. Phys. Chem. B 1997;101:10506–10517.
211. Cossi M, Scalmani G, Rega N, Barone V. J. Chem. Phys 2002;117:43–54.
212. Florián J, Warshel A. J. Phys. Chem. B 1999;103:10282–10288.
213. King G, Warshel A. J. Chem. Phys 1989;91:3647–3661.
214. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JJA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji M, Hada M, Ehara K, Toyota R, Fukuda J, Hasegawa M, Ishida T, Nakajima Y, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adacmo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski J, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas Ö, Malick DK, Rabuck AD, Clifford K, Cioslowki J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson BG, Chen W, Wong MW, Gonzalez C, Pople JA. GAUSSIAN 03 (Revision C.02). 2004
215. Adamo C, Barone V. J. Chem. Phys 1998;108:664–675.
216. Besler BH, Merz KM, Kollman PA. J. Comput. Chem 1990;11:431–439.
217. Haranczyk M, Gutowski M. J. Chem. Inf. Model 2007;47:686–694. [PubMed: 17381178]
218. Deisenhofer J, Steigemann W. Acta Crystallogr. sect. B 1975;31:238–250.
219. Wütrich K, Wagner G. J. Mol. Biol 1979;130:1–18. [PubMed: 38342]
220. Wolfenden R, Andersson L, Cullis PM, Southgate CB. Biochemistry 1981;20:849–855. [PubMed: 7213619]
221. Dao-pin S, Anderson DE, Baase WA, Dahlquist FW, Matthews BW. J. Mol. Biol 1991;1991:11521–11529.
222. Sakane SY, Yezdimer EM, Liu W, Barriocanal JA, Doren DJ, Wood RH. J. Chem. Phys 2000;113:2583–2593.
223. Boudaiffa B, Cloutier P, Hunting D, Huels M, Sanche L. Science 2000;287:1658–1660. [PubMed: 10698742]
224. Seidel CAM, Schulz A, Sauer MHM. J. Phys. Chem 1996;100:5541–5553.
225. Aflatooni K, Gallup GA, Burrow PD. J. Phys. Chem. A 1998;102:6205–6207.
226. Periquet V, Moreau A, Carles S, Schermann JP, Desfrancois C. J. Electron. Spectrosc. Relat. Phenom 2000;106:141–151.
227. Li X, Cai Z, Sevilla MD. J. Phys. Chem. A 2002;106:1596–1603.
228. Wesolowski SS, Leininger ML, Pentchew PN, Schaeder HFI. J. Am. Chem. Soc 2001;123:4023–4028. [PubMed: 11457153]
229. Dolgunitcheva O, Zakrewski VG, Ortiz JV. J. Phys. Chem. A 2001;105:8782–8786.
230. Tonzani S, Greene CH. J. Chem. Phys 2006;124:054312. [PubMed: 16468874]
231. Ptasinska S, Denifl S, Mroz B, Probst M, Grill V, Illenberger E, Scheier P, Mark TD. Bond selective dissociative electron attachment to thymine 2005;124:124302.
232. Abdoul-Carime H, Gohlke S, Illenberger E. Phys. Rev. Lett 2004;92:168103. [PubMed: 15169265]
233. Zakjevskii VV, King SJ, Dolgounitcheva O, Zakrewski VG, Ortiz JV. 2006;128:13350–13351.
234. Bachorz RA, Rak J, Gutowski M. Phys. Chem. Chem. Phys 2005;7:2116–2125.
235. Haranczyck M, Rak J, Gutowski M. J. Phys. Chem. A 2005;109:11495–15503. [PubMed: 16354040]
236. Bachorz RA, Klopper W, Gutowski M. J. Chem. Phys 2007;126:085101. [PubMed: 17343472]
237. Warshel A, Parson WW. Ann. Rev. Phys. Chem 1991;42:279–309. [PubMed: 1747189]
238. Parson WW, Warshel A. Q. Rev. Biophys 2001;34:563–679. [PubMed: 11852595]
239. Stephens PJ, Jollie DR, Warshel A. Chem. Rev 1996;96:2491–2514. [PubMed: 11848834]
240. Han WG, Lovell T, Noodleman L. Inorg. Chem 2002;41:205–218. [PubMed: 11800609]
241. Rustad JR, Rosso KM, Felmy AR. J. Chem. Phys 2004;120:7607–7615. [PubMed: 15267673]
242. Blumberger J, Sprik M. Theor. Chem. Acc 2006;115:113–126.
243. Churg AK, Warshel A. Biochemistry 1986;25:1675–1681. [PubMed: 3011070]
244. Jensen GM, Bunte SW, Warshel A, Goodin DB. J. Phys. Chem. B 1998;102:8221–8228.

245. Newton MD. Coord. Chem. Reviews 2003;238:167–185.
246. Cascella M, Magistrato A, Tavernelli I, Carloni P, Rothlisberger U. Proc. Natl. Acad. Sci. U. S. A 2006;103:19641–19646. [PubMed: 17179046]
247. Blumberger J, Klein ML. J. Am. Chem. Soc 2006;128:13854–13867. [PubMed: 17044714]
248. Warshel A, Creighton S, Parson WW. J. Phys. Chem 1988;92:2696.
249. Westheimer FH. Chem. Rev 1981;64:317–352.
250. Cleland WW, Hengge AC. Chem. Rev 2006;106:3252–3278. [PubMed: 16895327]
251. Vetter IR, Wittinghofer A. Q. Rev. Biophys 1999;32:1–56. [PubMed: 10800520]
252. Hengge A. Adv. Phys. Org. Chem 2005;40:49.
253. Aqvist J, Kolmodin K, Florian J, Warshel A. Chemistry and Biology 1999;6:R71–R80. [PubMed: 10074472]
254. Friedman JM, Freeman S, Knowles JR. J. Am. Chem. Soc 1988;110:1268–1275.
255. Kirby JA, Jencks WP. J Am. Chem. Soc 1965;87:3209–3216.
256. Kirby JA, Varvoglis AG. J. Am. Chem. Soc 1967;89:415–423.
257. Florian J, Aqvist J, Warshel A. J. Am. Chem. Soc 1998;120:11524–11525.
258. Florian J, Warshel A. J. Am. Chem. Soc 1997;119:4458–4472.
259. Mercero JM, Barrett P, Lam CW, Fowler JE, Ugalde JM, Pedersen LG. J. Comput. Chem 2000;21:43–51.
260. Liu Y, Gregersen A, Hengge A, York DM. Biochemistry 2006;45:10043–10053. [PubMed: 16906762]
261. Liu Y, Gregersen BA, Lopez X, York DM. J. Phys. Chem. B 2006;109:19987–20003. [PubMed: 16853584]
262. Rosta E, Kamerlin SCL, Warshel A. Biochemistry 2008;47:3725–3735. [PubMed: 18307312]
263. Kamerlin SCL, Wilkie J. Org. Biomol. Chem 2007;5:2098–2108. [PubMed: 17581653]
264. Kamerlin SCL, Florian J, Warshel A. Chem. Phys. Chem. 2008In Press
265. Kamerlin SCL, Williams NH, Warshel A. J. Org. Chem. 2008In Press
266. Lopez X, Dejaegere A, Leclerc F, York DM, Karplus M. J. Phys. Chem. B 2006;110:11525–11539. [PubMed: 16771429]
267. Barnes JA, Wilkie J, Williams IH. J. Chem. Soc. Farady Trans 1994;90:1709–1714.
268. Ba-Saif SA, Davis AM, Williams AJ. J. Org. Chem 1989;54
269. Radhakrishnan R, Schlick T. J. Am. Chem. Soc 2005;127:13245–13252. [PubMed: 16173754]
270. Radhakrishnan R, Schlick T. Biochem. and Biophys. Research. Commun 2006;350:521–529. [PubMed: 17022941]
271. Bojin MDST. J. Phys. Chem. B 2007;111:11244–11252. [PubMed: 17764165]
272. Burley SK, Alsmo S, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. Nature Genet 1999;23:151–157. [PubMed: 10508510]
273. Jones S, Thornton JM. Curr. Opin. Chem. Biol 2004;8:3–7. [PubMed: 15036149]
274. Whisstock JC, Lesk AM. Q. Rev. Biophys 2003;2003:307–340. [PubMed: 15029827]
275. Schaefer P, Riccardi D, Cui Q. J. Chem. Phys 2005;123:014905. [PubMed: 16035867]
276. Gao J, Alhambra C. J. Chem. Phys 1997;107:1212–1217.
277. Gregersen A, York DM. J. Phys. Chem. B 2005;109:536–556. [PubMed: 16851046]
278. Warshel A. J. Phys. Chem 1979;83:1640–1652.
279. Rashin AA. 1990;94:1725–1733.
280. Cramer CJ, Truhlar DG. J. Am. Chem. Soc 1991;113:8305–8311.
281. Jorgensen WL, Briggs JM. J. Am. Chem. Soc 1989;111:4190–4197.
282. You TJ, Bashford D. Biophys. J 1995;69:1721–1733. [PubMed: 8580316]
283. Yang A-S, Gunner MR, Sampogna R, Sharp K, Honig B. PROTEINS: Structure, Function, Genetics 1993;15:252–265.
284. Antosiewicz J, McCammon JA, Gilson MK. J. Mol. Biol 1994;238:415–436. [PubMed: 8176733]
285. Allewell NM, Oberoi H. Methods Enzymol 1991;202:3–19. [PubMed: 1784179]

286. Sham YY, Chu ZT, Warshel A. J. Phys. Chem. B 1997;101:4458–4472.
287. Bashford D, Karplus M. Biochemistry 1990;29:10219–10225. [PubMed: 2271649]
288. Ghosh N, Cui Q. J. Phys. Chem. B 2008;112:8387–8397. [PubMed: 18540669]
289. Kato M, Warshel A. J. Phys. Chem. B 2006;110:11566–11570. [PubMed: 16771433]
290. Fitch CA, Karp DA, Lee KK, Stites WE, Lattman EE, Garcia-Moreno EB. Biophys. J 2002;82:3289–3304. [PubMed: 12023252]
291. Denisov VP, Schlessman JL, Halle B, Garcia-Moreno EB. Biophys J 2004;87:3982–3994. [PubMed: 15377517]
292. Karp DA, Gittis AG, Stahley MR, Fitch CA, Stites WE, Garcia-Moreno EB. Biophys J 2007;92:2041–2053. [PubMed: 17172297]
293. Harms MJ, Schlessman JL, Chimenti MS, Sue GR, Damjanovic A, Garcia-Moreno EB. Protein Sci. 2008In Press
294. Sharma PK, Xiang Y, Kato M, Warshel A. Biochemistry 2005;44:11307–11314. [PubMed: 16114867]
295. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. Science 2000;289:920–930. [PubMed: 10937990]
296. Borman S. Chem. Eng. News 2004;82:29–30.
297. Katunin VI, Muth GW, Strobel SA, Wintermeyer W, Rodnina MV. Mol. Cell 2002;10:339–346. [PubMed: 12191479]
298. Sievers A, Beringer M, Rodnina MV, Wolfenden R. Proc. Natl. Ada. Sci. U. S. A 2004;101:7897–7901.
299. Gregory ST, Dahlberg AE. Nat. Struct. Mol. Biol 2004;11:586–587. [PubMed: 15221020]
300. Weniger JS, Parnell KM, Dorner S, Green R, Strobel SA. Nat. Struct. Mol. Biol 2004;11:1101–1106. [PubMed: 15475967]
301. Zhang L, Daiqian X, Dingguo X, Guo H. Chem. Commun 2007:1638–1640.
302. Liu Y, Gregersen BA, Hengge A, York DM. Biochemistry 2006:24708–24719.
303. Im W, Roux B. J. Chem. Phys 2001;115:4850–4861.
304. Hu C-H, Brinck T. J. Phys. Chem. A 1999;103:5379–5386.
305. Alden RG, Parson WW, Chu ZT, Warshel A. J. Am. Chem. Soc 1995;117:12284–12298.
306. Warshel, A.; Creighton, S. Microscopic free energy calculations of solvated macromolecules as a primary structure-function correlator and the MOLARIS program. Leiden: ESCOM; 1989.

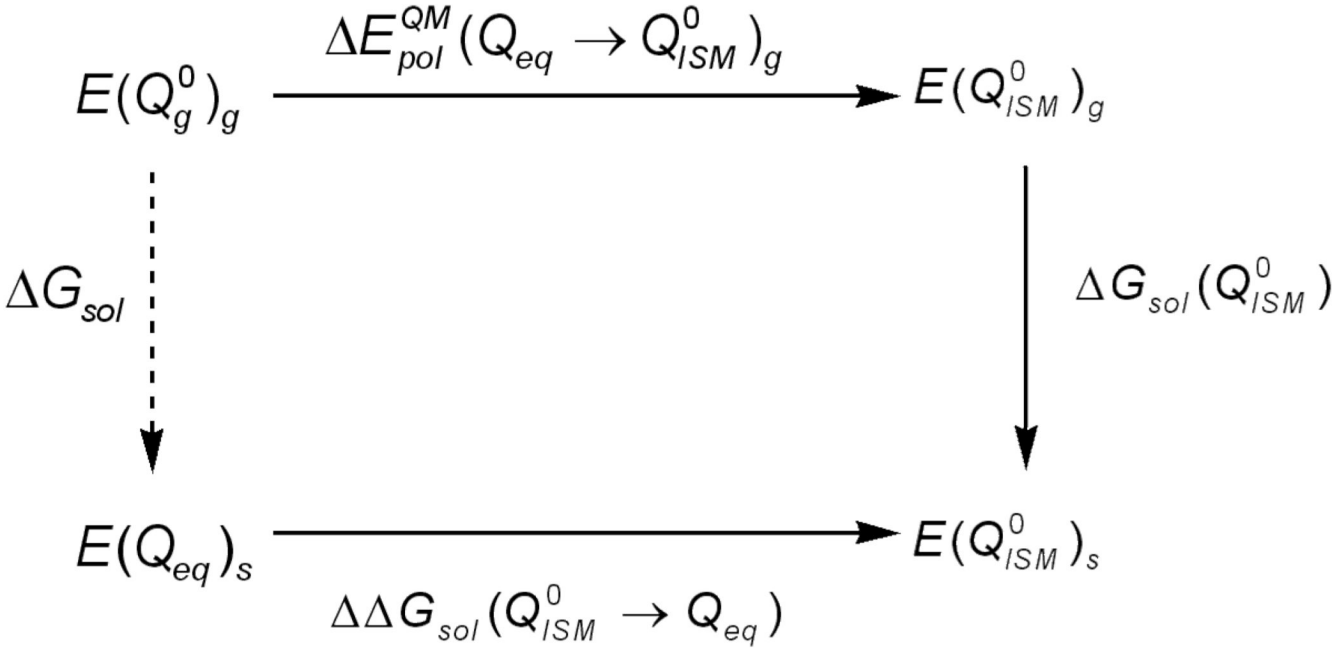


Figure 1.

Energy scheme for evaluating the solvation free energy. Here, Q^0_{ISM} is the charge distribution obtained by a given implicit solvation model, $\Delta G_{\mathrm{sol}}(Q^0_{\mathrm{ISM}})$ is evaluated by the classical FEP-AC approach, and all other terms are evaluated using a hybrid QM/MM approach.

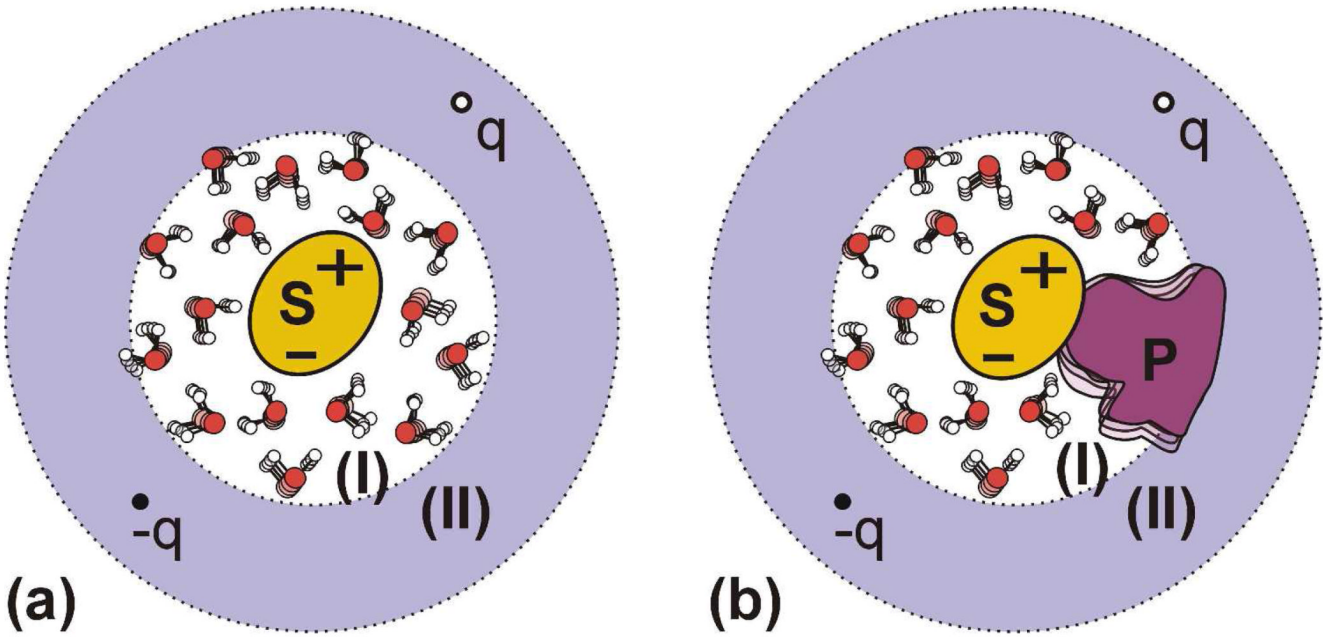


Figure 2.

Model for evaluating the average solute (S) charges for (a) a molecule in solution and (b) a protein sidechain. All atoms in Region I are represented explicitly, while Region II atoms are represented by two charges. In case (a), Region I only comprises explicit water molecules while in case (b) it also comprises all electroneutral groups within a pre-defined cut-off of the relevant protein sidechain. “P” designates the protein.

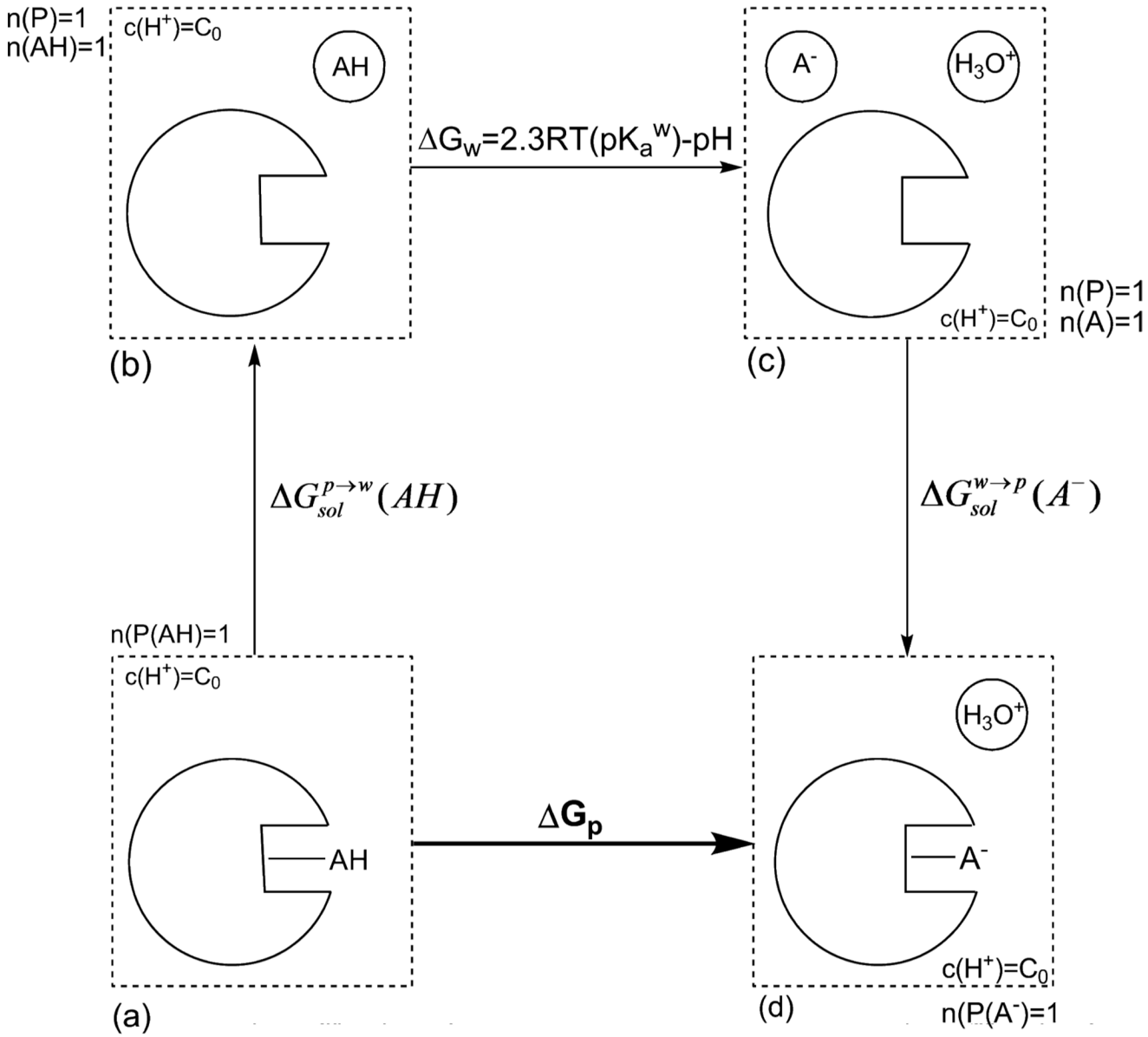


Figure 3.

A graphical description of the thermodynamic cycle for estimating the energetics of acid dissociation in a protein[152]. The ionisation process has been represented relative to the energetics and solvation of the A^- and AH species for the corresponding dissociation in water. Here, ΔG_p and ΔG_w represent the free energy of ionising an acid in a protein and water respectively, $\Delta G_{solv}^{p\rightarrow w}$ and $\Delta G_{solv}^{w\rightarrow p}$ represent the difference in solvation energy of the indicated species in protein (p) and water (w) respectively, R is the ideal gas constant, and AH and A^- represent the ionised an neutral forms of the acid respectively.

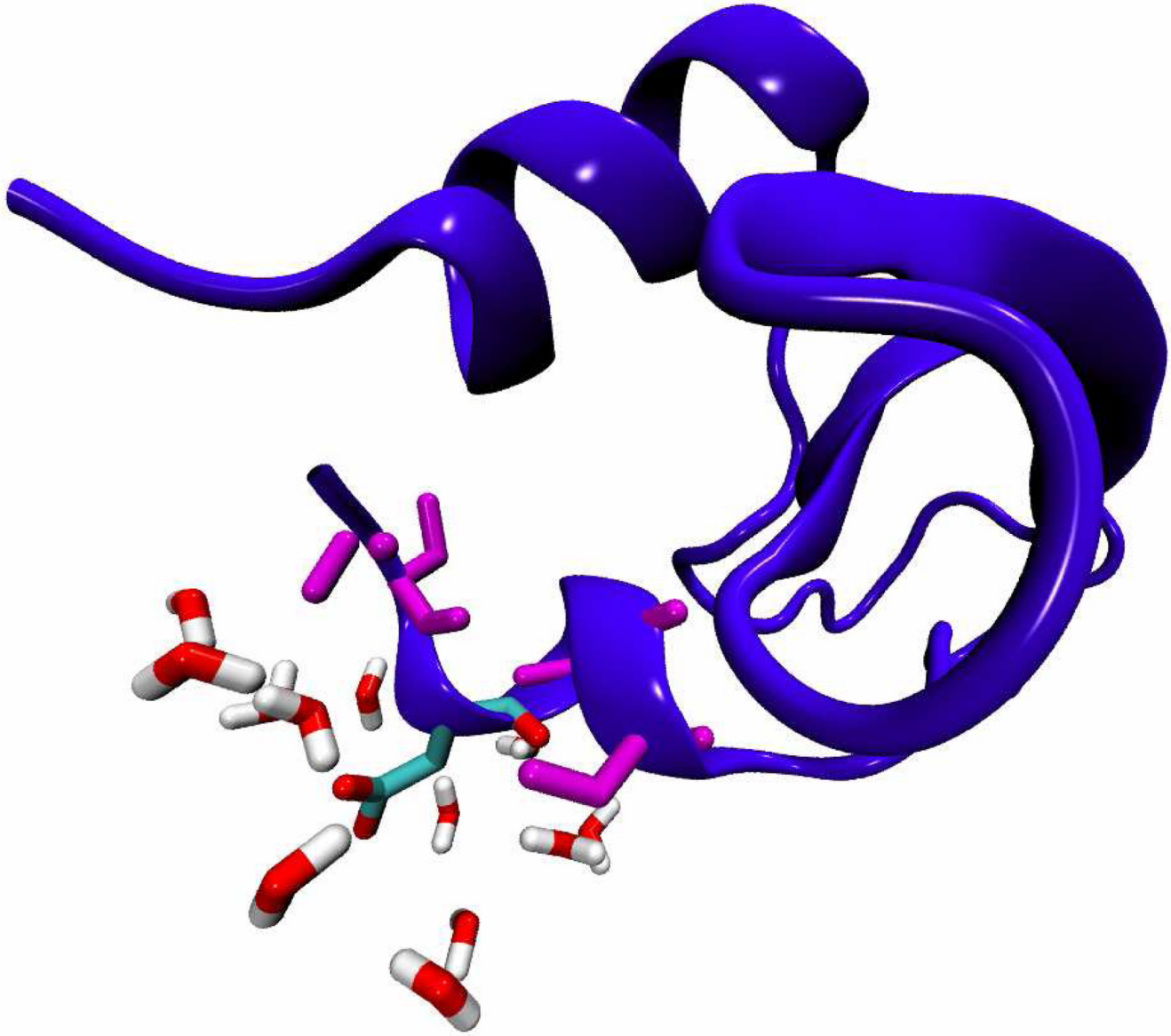
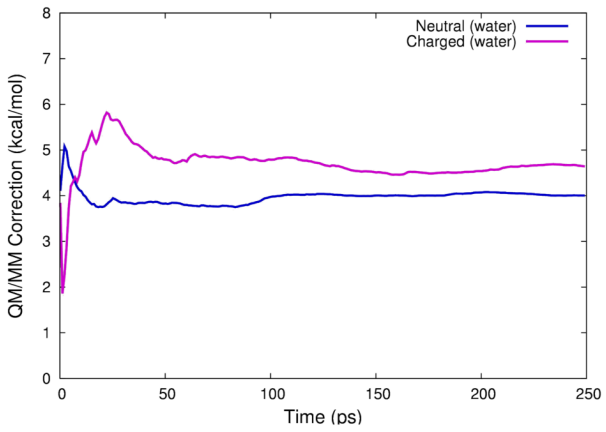
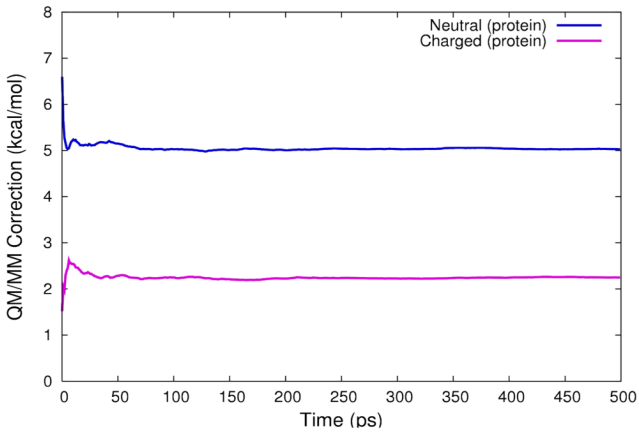


Figure 4.
The position of the Asp3 sidechain on the surface of the bovine pancreatic trypsin inhibitor (BPTI). The protein is shown in violet, and Asp3 is coloured by atom type. Also highlighted are all water molecules and electroneutral groups (shown in magenta) within 5Å of Asp3.



(a)



(b)

Figure 5.

The contribution to the total free energy of solvation from the LRA terms of Eq. 6. (a) shows the free contribution for acetate in water, and (b) shows contribution for the Asp3 sidechain of the bovine pancreatic trypsin inhibitor (BPTI). In both cases, the neutral form is shown in blue and the ionised form in magenta.

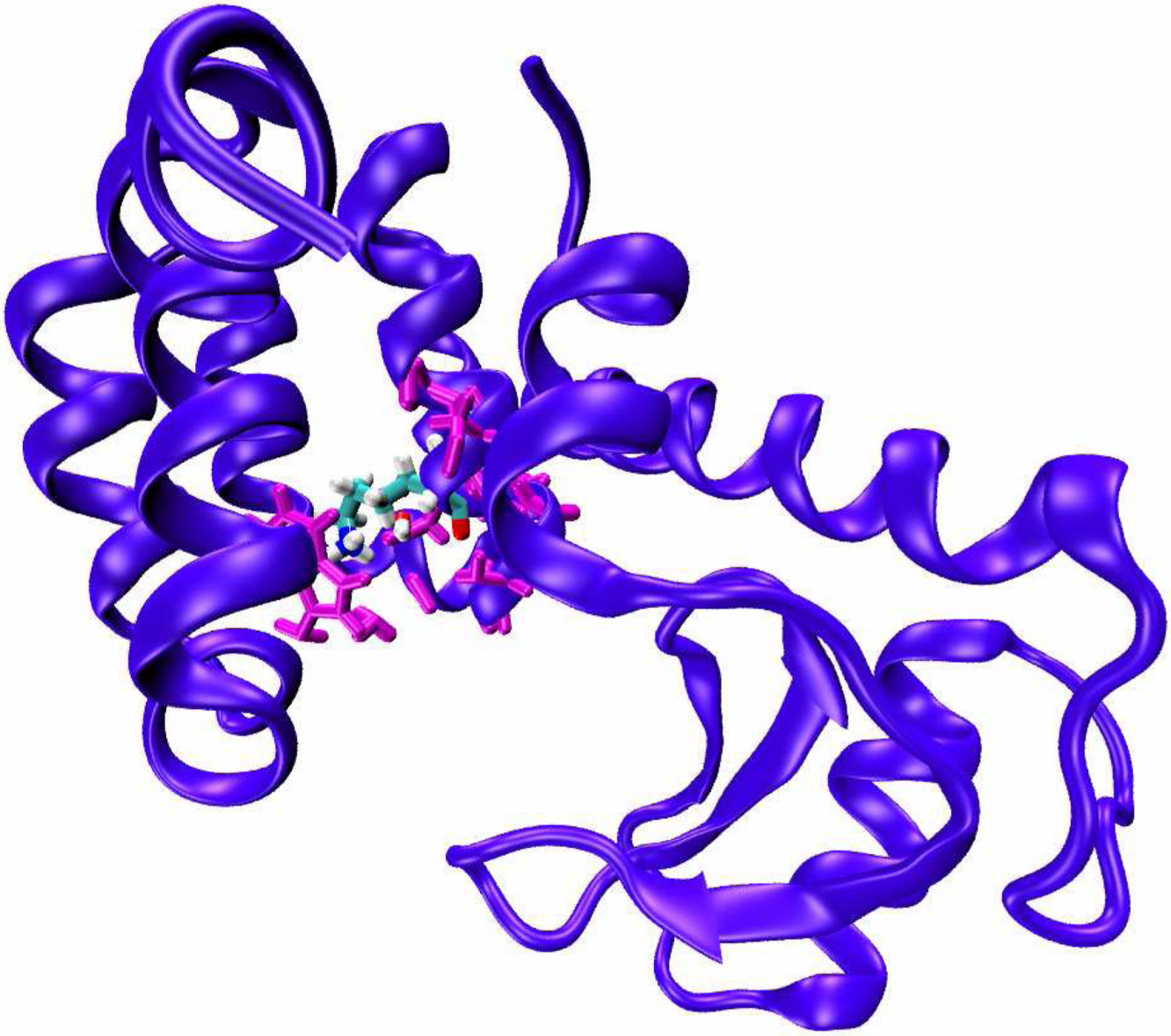


Figure 6.
The position of the Lys102 sidechain in the M102K lysozyme mutant. The protein is shown in violet, and Lys^{102} is coloured by atom type. Also highlighted are all water molecules and electroneutral groups (shown in magenta) within 5Å of Lys102.

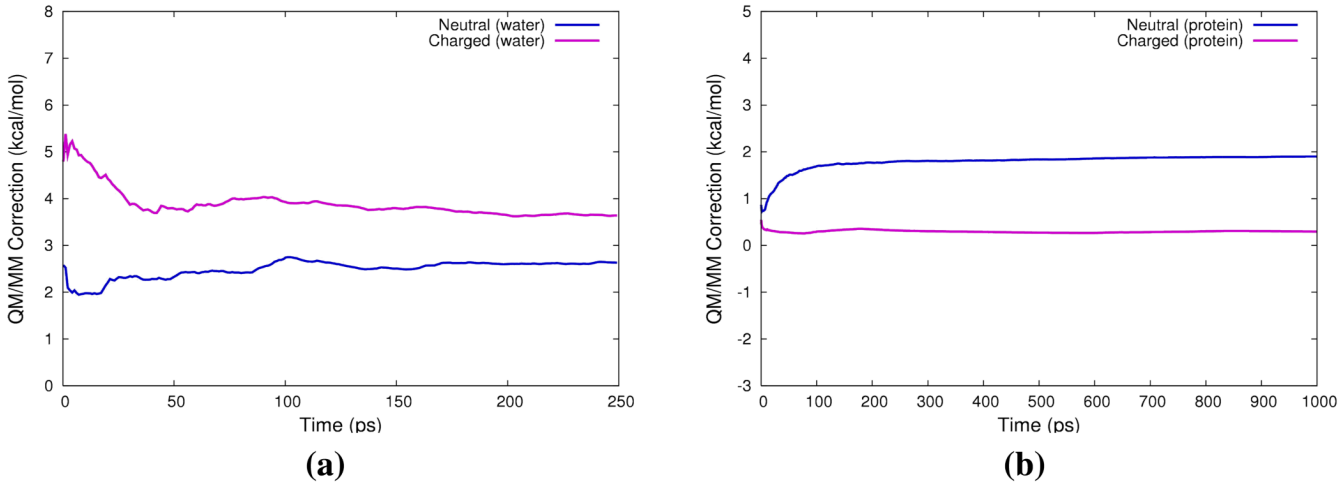


Figure 7.

The contribution to the total free energy of solvation from the LRA terms of Eq. 6. (a) shows the free contribution for methylamine in water, and (b) shows contribution for the Lys102 sidechain of the M102K mutated T4-lysozyme. In both cases, the neutral form is shown in blue and the ionised form in magenta.

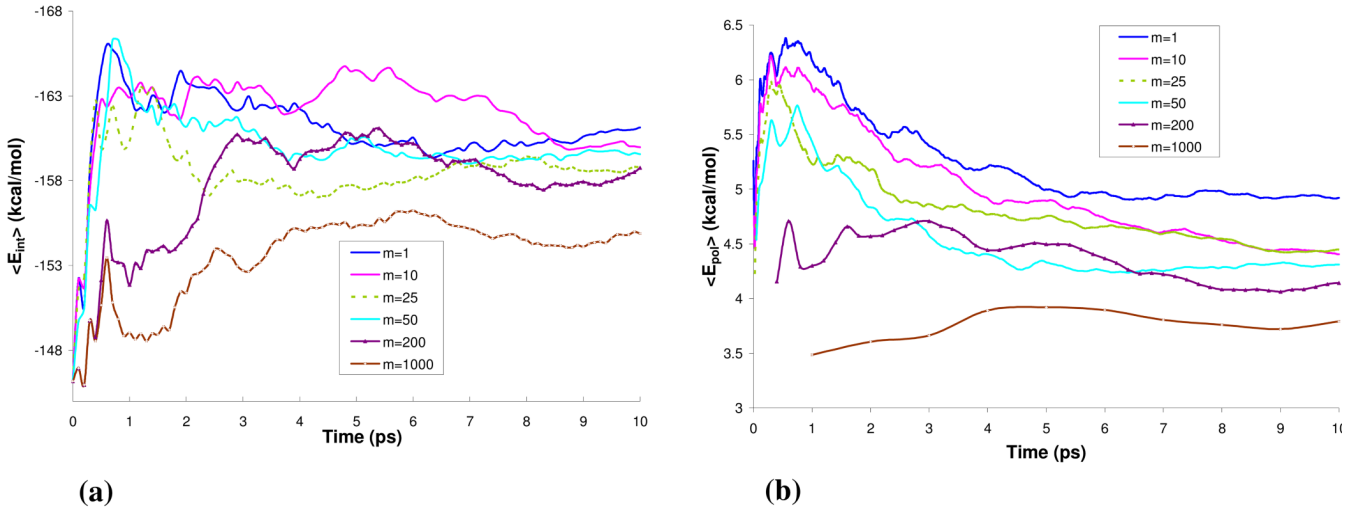


Figure 8.

Average (a) solute-solvent interaction energy and (b) solute polarization energy, along a 10 ps simulation of $HCOO^-$ solvated by water (using the same solvation model as that presented in Fig. 2b)[62]. The fact that the final sum of $\langle E_{int} \rangle$ and $\langle E_{pol} \rangle$ is independent of *m* cannot be realised from this figure, but rather from Fig. 9 of Ref.62.

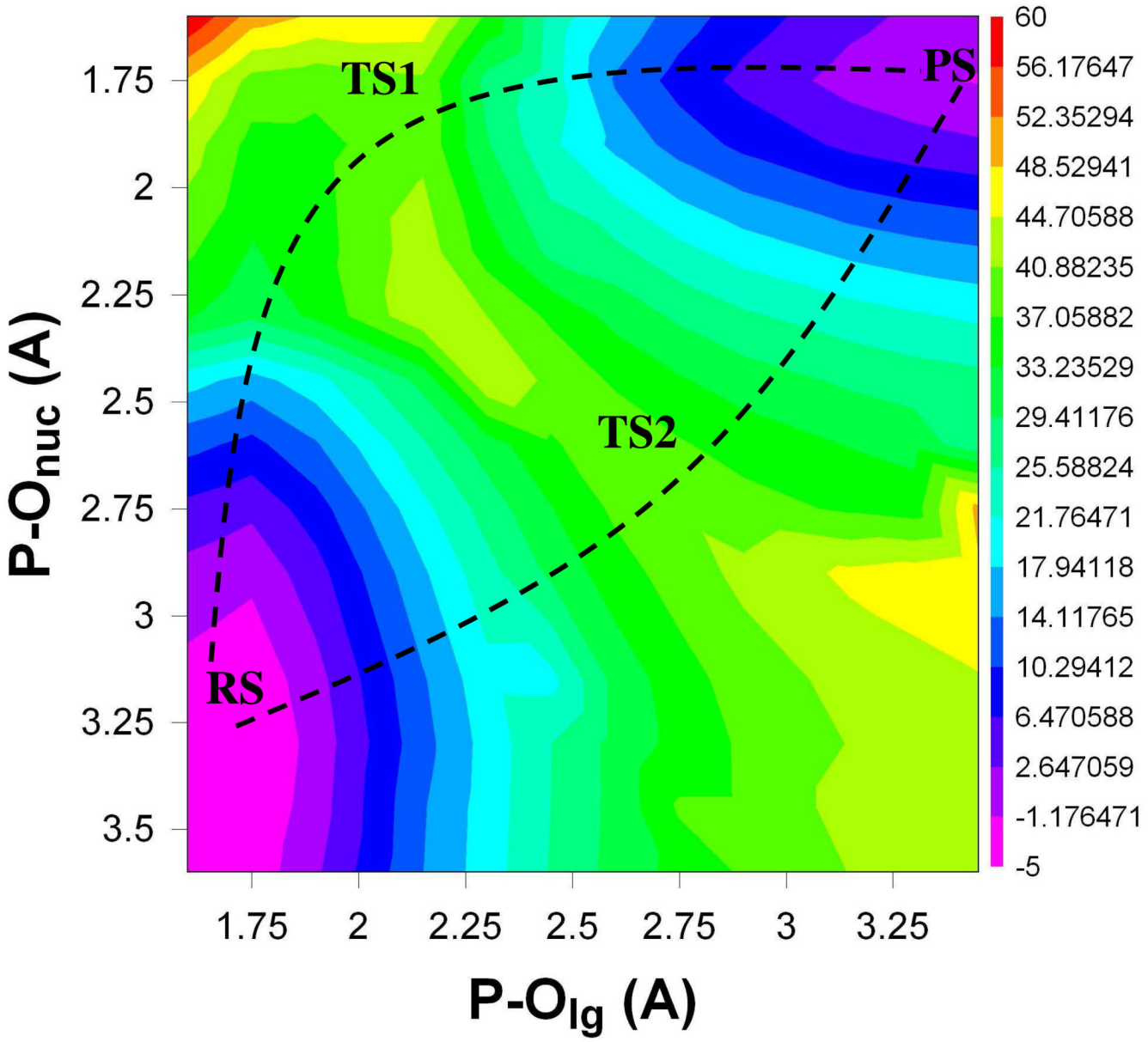


Figure 9.
Free energy surface for the hydrolysis of the methyl phosphate dianion ($MeOPO_3^{2-}$). Here, RS denotes the reactant state, PS the product state, and TS1 and TS2 associative and dissociative transition states respectively[264].

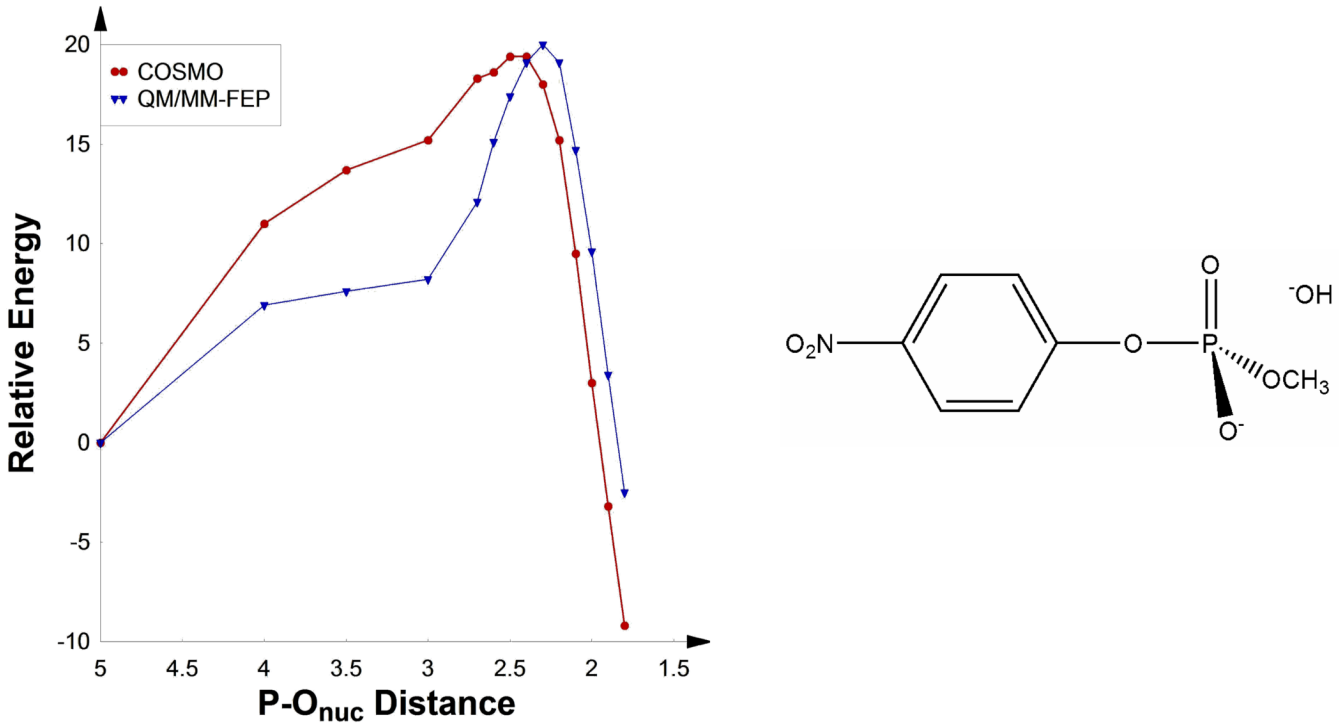


Figure 10.

A comparison of reaction profiles obtained by the COSMO solvation model (red) and the QM/MM-FEP approach presented in this work (blue) for hydroxide attack on the 4-nitro substituted methyl phenyl phosphate diester shown to the right, as a function of distance between the phosphorus atom being attacked and the oxygen atom of the incoming nucleophile (P-O_{nuc} distance). Distances are given in Å, and energies (relative to the diester and hydroxide at infinite separation) are given in kcal/mol.

Table 1

Summary of the solvation model used.

System	*Radius of Regions I and II in Å*		*Number of Explicit Water Molecules Solvating Different Systems*	
	Region I	**Region II**	**Neutral**	**Charged**
Acetate	16	16	560	562
Methylamine	16	16	561	561
BPTI	10	20	729	728
Lysozyme	10	20	317	317

Table 2

The contributions to the overall free energies of solvation of acetate in water and the Asp3 sidechain of BPTI.

System	$\Delta G_{sol}(Q = 0 \rightarrow Q^{0}_{COSMO})$	ΔG_{cav}	*QM/MM Correction*[a]	ΔG_{solv}	$\Delta\Delta G_{solv}$
Acetate (Neutral)	−12.4	1.3	3.9	−7.2	
Acetate (Charged)	−79.8	1.5	4.6	−73.7	
					−66.5
Asp3 (Neutral)	−13.0	--	2.9	−10.1	
Asp3 (Charged)	−79.0	--	4.7	−74.3	−64.2

[a]"QM/MM Correction" denotes the contribution from the LRA terms of Eq. 6. All energies given in kcal/mol.

Table 3

The contributions to the overall free energies of solvation of methylamine in water and the Lys102 sidechain of the M102K T4-lysozyme mutant.

System	$\Delta G_{sol}(Q = 0 \rightarrow Q^{0}_{COSMO})$	ΔG cav	*QM/MM Correction*[a]	ΔG solv	ΔΔG solv
Methylamine (Neutral)	−7.6	1.3	2.6	−3.7	
Methylamine (Charged)	−82.8	1.5	3.6	−77.7	
					−74.0
Lys102 (Neutral)	−2.2	--	2.9	0.7	
Lys102 (Charged)	−68.5	--	2.6	−65.9	−65.2

[a]"QM/MM Correction" denotes the contribution from the LRA terms of Eq. 6. All energies are given in kcal/mol.