JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Role of Deprotonation Free Energies in pK_a Prediction and Molecule Ranking

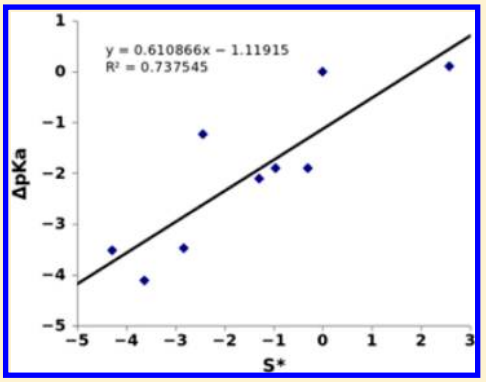
M. S. Bodnarchuk,*[,†] D. M. Heyes,[†] D. Dini,[†] S. Chahine,[‡] and S. Edwards[‡]

[†]Department of Mechanical Engineering, Imperial College London, Exhibition Road, London, SW7 2AZ, U.K.

[‡]BP Marine Limited, Marine Technology Centre, Whitchurch Hill, Pangbourne, RG8 7QR, U.K.

Supporting Information

ABSTRACT: A computationally efficient classical molecular simulation technique is derived for ranking the pK_a values of a set of chemically similar congeneric molecules in an implicit solvent model of water. This uses the deprotonation free energy of the titratable group in the gas and aqueous phases obtained by thermodynamic integration (TI). For a series of alcohols and acids a strong linear correlation is demonstrated between the experimental pK_a and the deprotonation free energy difference in the gas and liquid phases. These calculations also show that classical TI is more efficient than slow-growth TI in calculating deprotonation free energies for the series of molecules considered herein.



I. INTRODUCTION

The shape and structure of proteins and nucleotides can be sensitive to the pH of the system.[1−3] Changes in the pH and protonation state of ionizable groups within proteins can affect both the internal hydrogen bonding network and the external network between the protein and solvent molecules. This can lead to conformational changes which determine the physical and structural properties,[4−7] such as the catalysis rate in enzymes.[8] The pK_a of a drug molecule is an important physiochemical parameter which can influence its performance,[9] such as its dissolution into the bloodstream[10] and lipophilicity[11] which is important in determining the molecule's ability to cross the blood−brain barrier to target the central nervous system. Therefore, the ability to predict the pK_a's of potential drugs and similar types of molecules reliably, efficiently, and cost effectively is likely to become an increasingly important application of computational chemistry in the years to come. A computationally efficient free energy based approach for ranking the pK_a values of a series of chemically similar molecules is described in this work, which is an extension of existing methods for calculating the pK_a values of small molecules and proteins.

A number of computational chemistry techniques have been developed which can be used to predict the pK_a values of small molecules,[12−19] some of which use thermodynamic cycles, such as the one given in Figure 1. A quantum-mechanical (QM) representation is made of the solute, and a continuum model describes the solvent. Such approaches calculate the free energy of deprotonation of the molecule in the solvent by an indirect route, involving the gas-phase acidity, which can be found through high-level ab initio schemes, and the solvation free energies of the products and reactants obtained by classical simulation techniques.

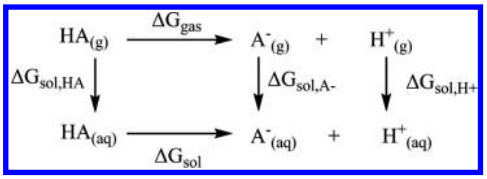


Figure 1. Thermodynamic cycle indicating the contributions to the calculation of pK_a of molecule HA through gaseous and solution processes.

The deprotonation free energy of a molecule HA in solution (the bottom line of the cycle given in Figure 1), ΔG_{sol}, formally defines the pK_a through

$$\Delta G_{sol} = 2.3RT\,pK_a = \Delta G_{gas} + \Delta G_{sol,A^-} + \Delta G_{sol,H^+} - \Delta G_{sol,HA} \quad (1)$$

pK_a prediction for small molecules using current computational methods suffers from a number of drawbacks.[16] The calculation of the deprotonation free energy in solution using a thermodynamic cycle requires knowledge of the free energy of hydration of a proton, $\Delta G_{sol,H^+}$, whose uncertainty leads to ±3 range for the final predicted pK_a value.[16] A second problem is the neglect of the contribution of structural water molecules around the solute molecule.[12,16]

Jorgensen and Briggs,[13,14] and Lim et al.,[15] showed that the effects of these problems can be minimized by considering the pK_a *difference* between two similar compounds. The difference in the gas-phase acidity between the two compounds can be

Received: October 21, 2013
Published: April 25, 2014

ACS Publications © 2014 American Chemical Society

dx.doi.org/10.1021/ct400914w | J. Chem. Theory Comput. 2014, 10, 2537−2545

calculated by high-level ab initio calculations. Classical molecular mechanics (MM) simulations are performed to calculate the alchemical free energy changes between the two compounds in their protonated and deprotonated states. This approach eliminates the need to include the solvation free energy of a proton, and gives the pK_a difference between two molecules in reasonable agreement with experiment.[13−15]

While various approaches have been developed for predicting the pK_a values of small molecules,[12−19] the value of amino acid pK_a in proteins has been found to be more challenging. The approach based on the thermodynamic cycle in Figure 1 is not practical for most proteins because of the large computational expense in calculating the QM components and the need to incorporate conformational flexibility around the titratable group of the molecule. Poisson−Boltzmann based methods have been used to calculate the pK_a values of proteins, although they can be extremely time-consuming for large systems.[20,21] Consequently, biomolecular simulations have until recently been performed with fixed protonation states for the titratable groups, identified as those forming the most likely hydrogen bonding network at a particular pH value.[22] While this approach is suitable for systems where major conformational changes are unlikely to occur between the two limiting states, it is less justifiable for more structurally mobile systems, especially where the pK_a of a critical residue is close to biological pH 7.[1] Such a representation of the protonated state can cause systematic errors in both the prediction of the acid−base equilibria and global conformational properties.

The protonation state of ionizable protein residues will in reality respond dynamically to an imposed pH. Constant-pH simulations allow the different protonation states of the residue(s) to be sampled as a function of pH to give a more realistic value for the pK_a of an ionizable residue in a protein.[23−34] This sampling is facilitated by using a reference compound which has the same titratable group as the ionizable residue in the protein and whose pK_a is known experimentally. Prior to constant-pH simulations the free energy of deprotonation of the reference compound is calculated using a free energy simulation, such as λ-dynamics, slow-growth, or classical thermodynamic integration (TI). Constant-pH simulations calculate the difference in the deprotonation free energies between the reference compound and the protein residue, which is proportional to the shift in pK_a between them. This is similar to the approach of Briggs et al.[13,14] and Lim et al.[15] The proton-donor bond dissociation free energy is assumed to be the same in the reference compound as the protein, which obviates the need to carry out computationally expensive QM calculations.[1]

The are two main categories of constant-pH simulation: the first uses the Metropolis Monte Carlo (MC) procedure to sample the protonation state of the residue, within the framework of a molecular dynamics (MD) simulation.[23,24,28,31−34] Periodically, a MC move is attempted to change the protonation state of the residue,[35] based upon the free energy change in the protein and the corresponding free energy change in the reference compound. The relative populations of the protonated and deprotonated states as a function of pH are used to define a deprotonation free energy and pK_a on fitting the numerical data to the Henderson−Hasselbalch (HH) formula.[33]

The second class of constant-pH simulation uses λ-dynamics[36] to sample the protonation state of the residue.[1,25−27,29,30] In these simulations λ is an intermolecular scaling parameter which describes the protonation state of the system, where λ = 0 represents the protonated form of the residue and λ = 1 represents the deprotonated form. Intermediate values of λ represent hybrid states. The parameter λ is treated as a dynamical variable which is driven by the corresponding potential energy change for the reference compound and the input value of the pH. A threshold value of λ is taken to define the protonated and deprotonated states as a function of λ, with the relative populations at the end of the simulation used to define the deprotonation free energy and, consequently, the pK_a.

In both the MC/MD and λ-dynamics approaches the deprotonation free energy of the reference compound is found prior to the simulation of the molecule with the unknown pK_a. In the first part of this work, a glutamic acid dipeptide was studied using both categories of constant-pH simulation, where the molecule was treated as both the reference molecule and the "protein". The deprotonation free energy and pK_a of the reference compound and the protein should be identical for such a system. The optimum method for calculating the reference deprotonation free energy is considered in section III, and then the relationship between the deprotonation free energy of a molecule and its pK_a is explored in section IV. It is shown in section IV that classically determined deprotonation free energies in implicit solvent and the gas phase enables the ranking of the pK_a values of a large number of compounds in aqueous solution to be made.

II. METHODOLOGY

a. Constant-pH Theory. The structure of the glutamic acid dipeptide used in this study is shown in Figure 2. As previously stated, the same molecule was used for both the reference (R) and protein (Prot) to explore the performance of different free energy methodologies.

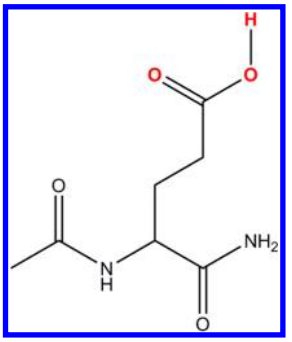


Figure 2. Chemical structure of the glutamic acid dipeptide modeled in this study. The titratable group atoms are shown in red.

The experimental deprotonation free energy of the reference compound, R(aq), is

$$\Delta G_{exp}(R) = \ln(10) k_B T(pK_{a\text{-}R} - pH) \quad (2)$$

where p$K_{a\text{-}R}$ is the pK_a of the model compound. At pH = pK_a the free energy difference between the protonated and deprotonated forms is zero, which corresponds to an equal population of the two states.

The constant-pH simulation method is based on the assumption that the difference in the experimentally determined deprotonation free energies of the reference compound and the protein is the same as that computed,[37] or equivalently

$$\Delta G_{exp}(Prot) - \Delta G_{exp}(R) = \Delta G_{MM}(Prot) - \Delta G_{MM}(R) \quad (3)$$

where ΔG_{exp} is the experimental free energy of deprotonation of either the reference (R) or the protein (Prot) system, and ΔG_{MM} is the calculated free energy of deprotonation from a classical MM free energy simulation. The titratable group in the protein can be regarded as the reference molecule which has been perturbed by the influence of the surrounding protein residues. The reference molecule deprotonation free energy, $\Delta G_{MM}(R)$, is calculated prior to the main constant-pH simulation, and in this study slow-growth TI and classical TI are used to obtain a potential of mean force for the gradual removal of the proton from the titratable group. The pK_a difference between the reference compound and the model protein can be calculated by TI.[38]

In constant-pH simulations, a pH term is introduced into the classical MM potential to calculate $\Delta G_{MM}(Prot)$. In the MC/MD approach implemented in AMBER, a Monte Carlo step attempts to add or remove a proton periodically during the MD simulation.[23,24,28,31−34] The free energy change for the protonation process of the protein in the simulation, ΔG_{Prot}, is

$$\Delta G_{Prot} = \ln(10) k_B T(pH - pK_{a\text{-}R}) + \Delta G_{elec} - \Delta G_{MM}(R) \quad (4)$$

where ΔG_{elec} is the electrostatic free energy upon the change in protonation state for the protein and $\Delta G_{MM}(R)$ is the corresponding free energy change for the reference compound.

The formula in eq 4 is used to sample the protonation state via a Metropolis Monte Carlo simulation. The probability of the titrated residue being deprotonated, P_D, is calculated during the simulation, plotted as a function of pH, and fitted to the Henderson−Hasselbalch (HH) formula:

$$P_D(pH) = 1/(1 + 10^{pK_a - pH}) \quad (5)$$

from which the pK_a value can be determined.

In the λ-dynamics approach,[1,25−27,29,30] a parameter, λ, dictates the degree to which the proton interacts with the system. The value of λ fluctuates between 0 (the protonated state) and 1 (the deprotonated state). The pH is introduced as a biasing potential in the potential energy function of the system. The biasing potential has two parts. The first term is $-\Delta G_{MM}(R)$, which reflects the free energy of deprotonation of the reference system as a function of λ, found through a prior classical MM free energy simulation. The second contribution incorporates the experimental deprotonation free energy of the reference compound, $\Delta G_{exp}(R)$ from eq 2, which captures the pH dependence of this free energy as a function of λ. The potential energy of the system, $U(\lambda)$, is

$$U(R) = (1 - \lambda)H_0 + \lambda H_1 + H_{env} - \lambda(\Delta G_{MM}(R) + \Delta G_{exp}(R)) \quad (6)$$

where H_{env} is the potential energy from the part of the force field which does not depend on λ. H_0 and H_1 are the potential energy parts of the Hamiltonians at $\lambda = 0$ and $\lambda = 1$, respectively, with the sum of these terms being equivalent to $\Delta G_{MM}(Prot)$. At the end of the simulation the distribution of system states between $\lambda = 0$ and $\lambda = 1$ is fitted to the Henderson−Hasselbalch formula to obtain the pK_a for the titratable group.

b. Simulation Details. Constant-pH simulations using λ-dynamics, and the accompanying reference free energy simulations, were performed for explicit solvent using a modified version of GROMACS v3.3.3.[1,39,40] The glutamic acid dipeptide was modeled using the OPLSA,[41] GROMOS96 53A6,[42] and Amber99 force fields.[43−45] The molecule was hydrated in a 5 nm × 5 nm × 5 nm cubic periodic MD box with approximately 4200 SPCE water molecules.[46] The initial configuration was subjected to 1000 steps of steepest descent minimization to remove unphysical overlaps of the molecules. The system was then allowed to come to equilibrium for 50 ps at 300 K in the NVT ensemble followed by 200 ps in the NPT ensemble using the Berendsen thermostat[47] for the bulk solvent with a coupling time of 0.1 ps. The dipeptide experienced a separate thermostat from the bulk solvent. The pressure in the simulation was set to 1 bar using the Berendsen barostat[47] with a coupling time of 1 ps. The particle mesh Ewald (PME) method[48] was used for the electrostatic interactions, adopting a grid spacing of 0.12 nm and an interpolation order of 4. Interactions between atoms within 1.0 nm were evaluated after each time step, with atom−atom interactions beyond 1.0 nm evaluated every five steps. The Lennard-Jones interaction truncation distance was 1.6 nm. A leapfrog equation of motion integrator was used with a time step of 2 fs. The bond distances and angles of water molecules were constrained with the SETTLE[49] algorithm and all other bonds were constrained using LINCS.[50]

The first step in the simulation procedure was to calculate the free energy of deprotonation, $\Delta G_{MM}(R)$, for the reference compound. Slow-growth TI was used to compute this free energy, with the nonbonded force field terms evolved from the protonated to deprotonated state at a rate $d\lambda/dt$ of 4×10^{-8} ps^{-1}. At the end of the simulation, the potential of mean force (PMF) was fitted to a third order polynomial in λ for use in eq 6. Classical TI simulations were also performed to calculate this free energy. Eleven equally spaced λ-windows between 0 and 1 were utilized, with each window simulated for 5 ns. Data were collected after the first nanosecond for equilibration. The value of $\Delta U/\Delta\lambda$ was calculated for each window, which was then used to generate a PMF fitted to a third order polynomial in λ.[1]

Constant-pH simulations were then performed according to the procedure and conditions of Donnini et al.,[1] for 5 ns in total, with data collection performed over the last 4.5 ns of the simulation. The criteria $\lambda < 0.1$ and $\lambda > 0.9$ were used to define the protonated and deprotonated states, respectively. A parabolic potential with a 3 kJ/mol energy barrier between the two limiting λ-states was used to bias the sampling toward the two end points.[1] The reference pK_a for the glutamic acid dipeptide was taken to be 4.4[43] and used in eq 2. The standard error was calculated from the averages over at least four simulations, commenced from independent initial states randomly chosen at the end of the equilibration process. The same structure was used for both the reference and constant-pH simulations to ensure self-consistency and that the predicted pK_a was equal to the reference pK_a.

The discrete MC/MD constant-pH simulations were performed using AMBER 11 and an implicit solvent representation. The glutamic acid dipeptide was modeled using the Amber99 force field,[43] with the generalized-Born model of Onufriev[51] (with the parameter igb=2 in AMBER) with a dielectric constant of 78.5 used to describe the implicit GB solvent model. The cutoff was set to 3.0 nm, with a salt concentration of 0.1 M. A Berendsen thermostat[47] was used with a time constant of 2.0 ps to maintain the temperature at 300 K. Bond distances were maintained constant using the SHAKE algorithm.[52] A time step of 2 fs was used for the MD, with a Monte Carlo titration attempted every five steps. The system was equilibrated for 200 ps, and then production data were collected for a further 5.0 ns.

The reference state free energies for the glutamic acid dipeptide and the congeneric molecules were obtained using classical TI within the AMBER package. Eleven equally spaced λ-windows between 0 and 1 were used, with each window simulated for 5 ns, with data collected for the last 4 ns of the simulation. The simulation parameters in these and the constant-pH simulations were the same. The congeneric molecules were modeled using the GAFF force field,[53] with atomic charges calculated using the AM1-BCC charge model.[54]

III. CONSTANT-PH SIMULATIONS

a. Choice of TI Methodology To Calculate ΔG_{MM}(R). Slow-growth TI and classical TI were used to calculate the value of ΔG_{MM}(R) for the capped glutamic acid dipeptide in explicit solvent. In principle the two methods should give identical results, and as shown in Figure 3 the PMF profiles for the two methods with the same box size are in good agreement with one another.

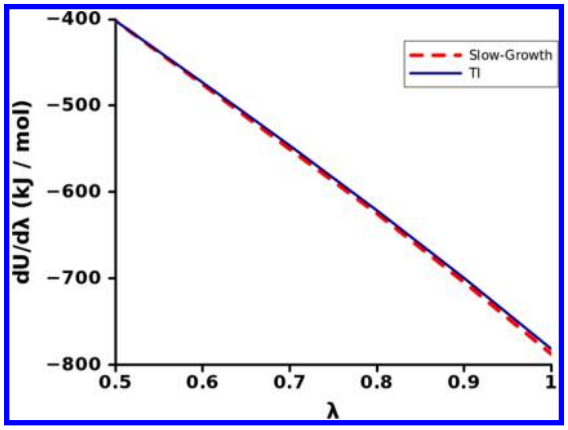


Figure 3. Reference-state PMF for the glutamic acid dipeptide, in the region $0.5 < \lambda < 1.0$, using slow-growth TI and classical TI. The OPLS force field was used. For the slow-growth TI, a rate of $d\lambda/dt = 4 \times 10^{-8}$ ps^{-1} was employed.

Although both methods give similar PMF profiles, the classical TI approach is favored for several reasons. First, each λ-value in classical TI can be carried out in parallel, making the method more efficient than slow-growth TI, which simulates each increment of λ sequentially in chronological order. Second, the slow-growth method can give poor convergence if the evolution rate is too fast. Simulations were initially performed using a λ-evolution rate of $d\lambda/dt = 4 \times 10^{-7}$ ps^{-1}, which showed a free energy profile that was increasingly different from Figure 3 around the $\lambda = 1$ end point. A "lag"[55] in the water structure around the end points of the simulation produced a wider range of end-point water-residue states than the classical TI method. The radial distribution functions demonstrating this behavior can be found in the Supporting Information. The classical TI method is preferred as the optimum λ evolution rate is not known a priori.

b. Constant-pH Simulations. Classical TI was chosen to calculate the deprotonation free energy, ΔG_{MM}(R), for the reference molecule in this work. Constant-pH simulations were performed on the dipeptide protein to check that the predicted pK_a from the simulation was the same as the experimental reference pK_a of 4.4, which indicates that the free energy calculated using λ-dynamics is the same as ΔG_{MM}(R), and that the phase space explored in the reference calculation and the main constant-pH simulation is essentially the same. Any discrepancy in the predicted pK_a value indicates a potential sampling issue in either simulation method.

Figure 4 shows the titration curves derived from λ-dynamics for the capped glutamic acid dipeptide molecule in explicit solvent using three different force fields.

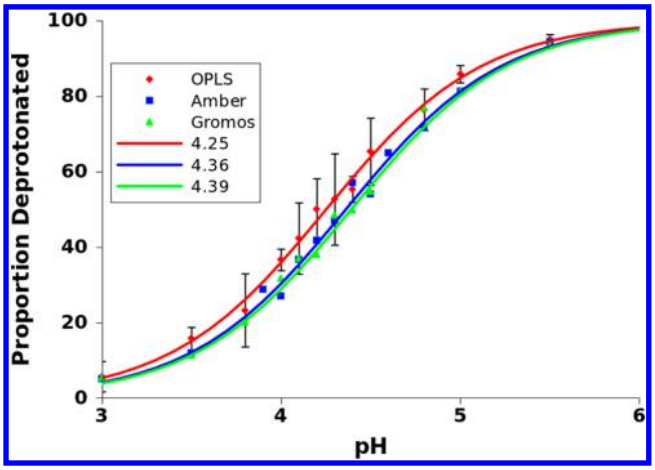


Figure 4. Percentage of the deprotonated molecules for the glutamic acid dipeptide, using the OPLS, Amber, and GROMOS53A6 force fields. Classical TI was used to calculate the value of ΔG_{MM}(R). Standard error bars given on the figure are calculated from at least four simulations, with only those for the OPLS force field shown for clarity. The continuous curves are the Henderson−Hasselbalch functions fitted to the simulation data.

The computed pK_a values for the glutamic acid dipeptide from the HH fit are 4.25 ± 0.03, 4.36 ± 0.04, and 4.39 ± 0.04 for the OPLS, Amber, and GROMOS force fields, respectively, which agree almost within statistical uncertainty with the experimental input value of 4.4, The value of ΔG_{MM}(R) is therefore the same as the free energy obtained through λ-dynamics, ΔG_{MM}(Prot). The slow-growth method, using a rate of $d\lambda/dt = 4 \times 10^{-7}$ ps^{-1}, was found to result in a significant pK_a offset in the main constant-pH simulation, and therefore the classical TI method is recommended.

The corresponding constant-pH simulations for the dipeptide were also performed in implicit solvent using the discrete MD/MC approach in AMBER.[56] The value of ΔG_{MM}(R) was calculated using classical TI. Figure 5 presents the HH fitted titration profile obtained from the classical TI deprotonation free energies. The structure and partial charges used were the same as in the previously described GROMACS simulations, and the fitted HH, pK_a = 4.40 ± 0.04, is in excellent agreement with the experimental input of 4.4.

The possibility of solvent rearrangement around the titratable group is a potential issue for the calculation of ΔG_{MM} in explicit solvent, and can require long simulation times to achieve adequate convergence. Corresponding implicit solvent calculations do not require significant equilibration since the solvent (dielectric) response to a change in charge is instantaneous[56] and are approximately 2 orders of magnitude faster to carry out. The HH curves obtained from the implicit solvent model are in excellent agreement with the explicit solvent cases, and also take approximately 2 orders of magnitude less CPU time to obtain a full titration curve. The implicit solvent approach has therefore a number of advantages for systems which are relatively small or do not have key structural water molecules near the titratable group (in which case, explicit solvent should be used for the constant-

dx.doi.org/10.1021/ct400914w | J. Chem. Theory Comput. 2014, 10, 2537−2545

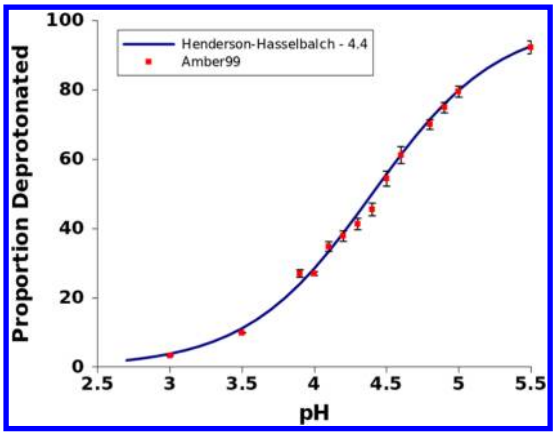


Figure 5. Percentage of deprotonated state for the glutamic acid dipeptide in implicit solvent using the Amber99 force field. The HH continuous curve on the figure is based on a best fit to the simulation data value of $pK_a = 4.4 \pm 0.04$. Standard errors given on the figure are derived from at least three independent simulations at the same pH value.

pH simulations). In section IV it is shown how these methodological procedures can be applied and extended to rank efficiently the pK_a values of a series of chemically similar molecules containing the same titratable group.

IV. MOLECULE pK_A RANKING

The free energy of the deprotonation reaction of a molecule in solution, ΔG_{sol}, the bottom line of the cycle given in Figure 1, is directly related to the pK_a of the molecule through eq 1. The bond-breaking component of ΔG_{sol} is quantum mechanical in nature[16] and therefore cannot be determined by a classical simulation technique. Also the free energy of solvation of the proton is not known precisely, so these issues need to be dealt with. The effects of these two terms can be minimized if the difference in the pK_a values between two compounds, HA and HB, is considered. Within a purely MM derived framework the difference in the two pK_a values is

$$\Delta\Delta G_{sol}^{QM} = 2.3RT\Delta pK_a = \Delta\Delta G_{gas}^{QM} + \Delta\Delta G_{sol,B/A^-}^{MM} - \Delta\Delta G_{sol,HB/HA}^{MM} \quad (7)$$

where $\Delta pK_a = pK_a(HB) - pK_a(HA)$, and $\Delta\Delta G = \Delta G(HB) - \Delta G(HA)$ is the difference in the free energy term between molecules HB and HA (taken later as the reference).[13−15] The three terms on the right-hand side of eq 7 represent the gas-phase deprotonation (which is quantum mechanical in nature), ion, and molecule solvation free energies, respectively. The solvation free energy of the gaseous proton, $\Delta G_{sol,H^+}$, cancels out in eq 7, by assuming it to be the same for HB and HA. Briggs et al.[13,14] and Lim et al.[15] used ab initio computational chemistry calculations to determine the gas-phase acidity difference, $\Delta\Delta G_{gas}$, between HB and HA. The solvation free energy difference between the protonated and deprotonated forms of the compounds were found here through MM-based alchemical transformations.

The thermodynamic cycle in Figure 6 shows that the MM solvation free energy difference for HA, $\Delta G_{sol,A^-} - \Delta G_{sol,HA}$, is equal to the MM-derived deprotonation free energies of HA in solvent and vacuum, $\Delta G_{dep\text{-}sol} - \Delta G_{dep\text{-}gas}$. Using this identity, eq 7 can be rewritten as

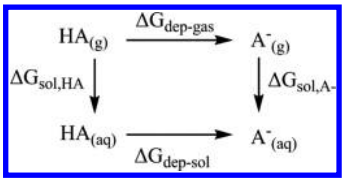


Figure 6. Thermodynamic cycle for the MM deprotonation reactions for the molecule HA in gaseous and aqueous phases.

$$\Delta\Delta G_{sol}^{QM} = 2.3RT\Delta pK_a = \Delta\Delta G_{gas}^{QM} + \Delta\Delta G_{dep,sol}^{MM} - \Delta\Delta G_{dep,gas}^{MM} \quad (8)$$

adopting the same $\Delta\Delta G$ notation scheme. As an approximation, the gas-phase acidity difference between the two molecules, $\Delta\Delta G_{gas}$, can be set to zero for structurally similar molecules, a not unreasonable approximation, so eq 8 then becomes

$$\Delta\Delta G_{sol}^{QM} \approx \Delta\Delta G_{dep,sol}^{MM} - \Delta\Delta G_{dep,gas}^{MM} \quad (9)$$

Dropping the MM superscript, from eq 9, the quantity, S^*

$$S^* = \frac{\Delta\Delta G_{dep,sol} - \Delta\Delta G_{dep,gas}}{2.303RT} \quad (10)$$

should therefore be equal to ΔpK_a to a first approximation.

Figure 7 shows the two sets of congeneric molecules used to investigate this MM approach. For each congeneric set of molecules a reference structure (labeled **C1** in both cases) is chosen. Two sets of simulations were then performed on each molecule in the aqueous and gas phases to calculate the values of $\Delta G_{dep,sol}$ and $\Delta G_{dep,gas}$.

An implicit solvent representation (relative permittivity, $\varepsilon = 78.5$) was chosen to model the solvent in this study. As previously shown, implicit solvent calculations are orders of magnitude more efficient than those using explicit solvent, making them more suitable for treating a large data set. The congeners within a series are of similar size but have slightly different intramolecular chemical environments around the titratable group, whose effect on the pK_a we consider should be captured quite well using the implicit solvent TI model. Figures 8 and 9 present the experimental pK_a differences plotted against the MM free energy parameter, S^*, found using eq 10.

Figures 8 and 9 show that there is a reasonably good anticorrelation between the pK_a shifts compared to the chosen reference molecule and the parameter S^* defined in eq 10. The predictive index[58] for the alcohol and acid subsets are −0.98 and −0.89 respectively, but the slopes deviate from the expected minus unity value (−0.28 ± 0.1 in both cases). The deficiency inherent in the derivation of S^* in eq 10 is because the difference between the deprotonation gas-phase free energy of the two molecules was set to zero, which can be corrected for in an approximate way. First this contribution is reintroduced:

$$\Delta\Delta G_{sol}/2.3RT = \Delta pK_a = \Delta\Delta G_{gas}/2.3RT + S^* \quad (11)$$

where $\Delta\Delta G_{gas}$ is the relative gas-phase acidity free energy difference between HB and HA.

Figures 8 and 9 show that the pK_a of the reference congener, HA, is typically the highest of the series (therefore ΔpK_a is typically negative). The value of $\Delta\Delta G_{gas}$ is thus anticipated to be negative for most of the congener pairs, and will reduce the currently defined values of S^*. Assuming that the gas-phase acidity free energy difference between two compounds is

dx.doi.org/10.1021/ct400914w | J. Chem. Theory Comput. 2014, 10, 2537−2545

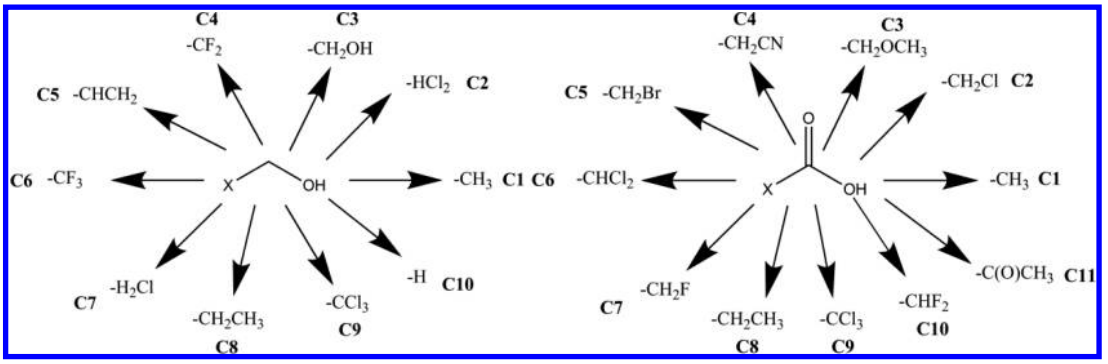


Figure 7. Two sets of congeneric molecules; alcohols on the left and acids on the right. Compound reference numbers shown in bold are also used to refer to the molecules in subsequent figures. The experimentally determined pK_a values of these molecules are given in the Supporting Information.[57]

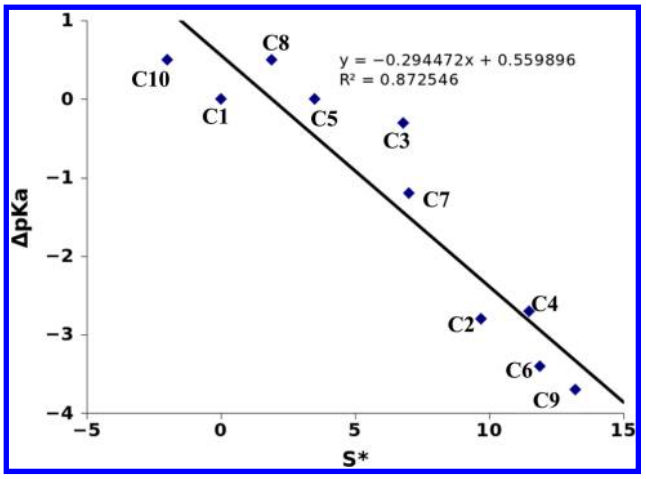


Figure 8. Predicted pK_a shifts from the reference, S^*, defined in eq 10, for the alcohol series plotted against the experimental pK_a shift from the reference molecule value, ΔpK_a. All shifts are relative to the reference molecule **C1**. The equation in the top right is the linear regression equation, $y = mx + c$, together with the calculated regression coefficient, R^2.

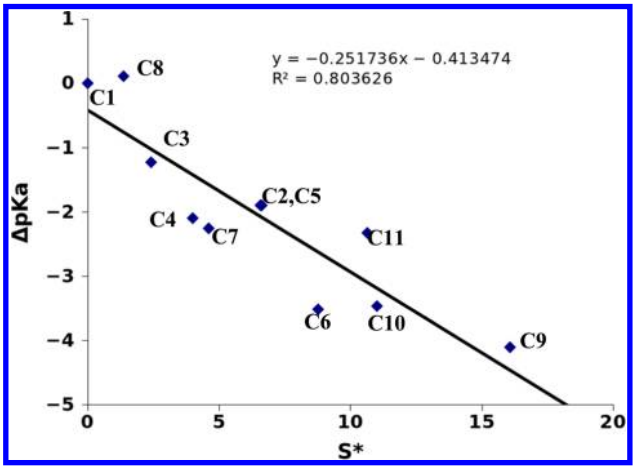


Figure 9. As for Figure 8 except that the acid series of molecules given in Figure 8 is considered.

proportional to the difference in the pK_a, to a first approximation, with a proportionality constant α, then

$$\Delta pK_a = \alpha \Delta pK_a + S^* \quad (12)$$

Upon rearrangement of eq 12

$$\Delta pK_a = S^*/(1 - \alpha) \quad (13)$$

for which $\alpha \approx 4$ complies with the trends in Figures 9 and 10. This correction does not affect the ranking of the molecules, just the slope of the linear regression fit to the data.

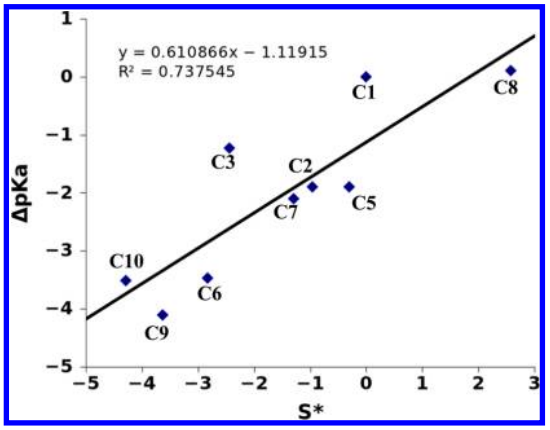


Figure 10. As for Figure 9 except the internal dielectric has the value ε = 1.5. The predictive index for this data set is 0.98. Compounds **C4** and **C11** are not shown since they exhibited behavior which was uncommon to the rest of the data set upon a change in the internal dielectric.

An alternate way of ensuring the slope is closer to unity is to adjust the internal dielectric constant of the solute. The molecules in the series with a weaker O−H bond strength are expected to be more polar than those with stronger O−H bonds, and this polarization effect could be incorporated through modifying the dielectric interior. Hou has shown that relative protein−ligand affinities calculated using MM/GBSA can be affected by the choice of this parameter, with some protein−ligand complexes exhibiting improved results when a dielectric of ε = 4 was chosen instead of 1 or 2.[59] To investigate this possibility, the S^* values were recalculated for the acetic acid subset using a dielectric constant of ε = 1.5, with the experimental pK_a differences plotted against the MM free energy parameter, S^*, shown in Figure 10.

Figure 10 shows that there is good quantitative agreement between the S^* values and the experimental pK_a shifts on changing the value of the internal dielectric constant. The slightly increased polar interior results in the deprotonation free energies performed in solvent becoming more negative for molecules which are expected to have weaker O−H bond energies (**C6**, **C9**, **C10**), while they become more positive for those with a stronger O−H bond energy (**C1**, **C8**). This results in a change in the sign

dx.doi.org/10.1021/ct400914w | J. Chem. Theory Comput. 2014, 10, 2537−2545

of the slope compared to the data presented in Figure 9. The incorporation of electronic polarizability could be considered to be equivalent to the α correction proposed above. A value of $\varepsilon = 1.5$ gives the best agreement with experiment for this type of molecule data set. The new value of the dielectric constant *significantly* increases the deprotonation free energy of **C4**, but only causes a small change for **C11**, molecules which become "outliers" and were omitted from the above analysis.

V. CONCLUSIONS

A purely classical molecular modeling method for placing the pK_a values of a chemically similar set of molecules in approximately the correct order is proposed and verified. This is based on the difference in the free energy of deprotonation of each molecule in the gas and liquid states, incorporated in the parameter S^*, which is predicted to be approximately proportional to the experimental pK_a value relative to a reference molecule. The magnitude and sign of the slope of experimental pK_a plotted against S^* is shown however to be sensitive to the value of the internal relative permittivity.

While there are a number of tools available for the rapid prediction and screening of pK_a values for small organic molecules, they are typically based on either empirical correlations, and hence limited to experimental data on known compounds, or simple force-field-based approaches.[60] The proposed approach is extremely fast (a set of 10 molecules can be easily studied in 1 h on a typical high-end PC) yet is rigorously founded on the relevant thermodynamic cycles. Another advantage of this predictive method is that the simulations can in principle be performed in the same way for any solvent or a mixture of solvents where the dielectric constant is known, but the experimental pK_a data may be scarce.

The classical TI method is shown in practice to be more efficient than slow-growth TI in calculating the deprotonation free energy in explicit solvent. For a typical or routine evolution rate of the deprotonation parameter, λ, in slow-growth TI the molecular conformation and solvation structure can lag behind the thermodynamically optimum state at that value of λ which biases the computed value of the free energy change. For systems where little or no solvent reorganization occurs between the two end points, then slow-growth TI could still be employed, although a (small) value of $d\lambda/dt < 4 \times 10^{-8}$ ps^{-1} would still be recommended to ensure all contributions to the free energy are completely accounted for (this rate will be system specific). Consequently, we recommend that, in order to ensure adequate phase space sampling with a minimum of prebenchmarking, slow-growth TI should be avoided for the majority of systems and classical TI used to calculate the PMF. In addition, if major solvent reorganization is anticipated through a large pH range, or if explicit solvent molecules are known to interact with the reference compound or titratable group,[61,62] advanced sampling methods such as replica exchange thermodynamic integration[63,64] or parallel tempering[65,66] may be required to ensure that the reference-state PMF incorporates all of the contributions from the Hamiltonian satisfactorily, such as the contribution from the solvent molecules around the titratable moiety.

The quality of the pK_a-computed free energy correlation computed here may deteriorate if there is conjugation (e.g., a benzene ring) between the titratable group and the modifying substituent as polarization and bond-strength effects will become more dominant. Computing the relative effects of ortho, meta, and para positional substitution in the same series would also pose a challenge, and then each position of substitution on the benzene ring should be treated as a separate series. To conclude, the ability to predict the relative values and order of the pK_a values of a series of small molecules in different solvents by classical molecular simulation techniques should provide a useful complementary tool to experiment in a number of practical applications.

■ ASSOCIATED CONTENT

Supporting Information

Radial distribution functions using a slow-growth rate of $d\lambda/dt = 4 \times 10^{-7}$ ps^{-1}. Experimental pK_a values for the two congeneric series of molecules. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: m.bodnarchuk@imperial.ac.uk. Tel.: +44 (0)20 7594 7242.

Funding

M.S.B., D.M.H., and D.D. thank BP Marine Ltd., Pangbourne, U.K., for funding this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

M.S.B., D.M.H., and D.D. thank Imperial College for providing additional computational resources.

■ REFERENCES

(1) Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmüller, H. Constant pH molecular dynamics in explicit solvent with λ-dynamics. *J. Chem. Theory Comput. Chem.* **2011**, *7* (6), 1962−1978.

(2) Farrell, D.; Miranda, E. S.; Webb, H.; Georgi, N.; Crowley, P. B.; McIntosh, L. P.; Nielsen, J. E. Titration_DB: Storage and analysis of NMR-monitored protein pH titration curves. *Proteins: Struct. Funct. Bioinf.* **2010**, *78* (4), 843−857.

(3) Goh, G.; Knight, J.; Brooks, C. Constant pH Molecular Dynamics Simulations of Nucleic Acids in Explicit Solvent. *J. Chem. Theory Comput.* **2011**, *8* (1), 36−46.

(4) Kozachkov, L.; Padan, E. Site-directed tryptophan fluorescence reveals two essential conformational changes in the Na^+/H^+ antiporter NhaA. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (38), 15769−15774.

(5) Lorieau, J. L.; Louis, J. M.; Schwieters, C. D.; Bax, A. pH-triggered, activated-state conformations of the influenza hemagglutinin fusion peptide revealed by NMR. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (49), 19994−19999.

(6) Bhat, V.; Kurouski, D.; Olenick, M. B.; McDonald, C. B.; Mikles, D. C.; Deegan, B. J.; Seldeen, K. L.; Lednev, I. K.; Farooq, A. Acidic pH promotes oligomerization and membrane insertion of the BclXL apoptotic repressor. *Arch. Biochem. Biophys.* **2012**, *528* (1), 32−44.

(7) Tournaire-Roux, C.; Sutka, M.; Javot, H.; Gout, E.; Gerbeau, P.; Luu, D.-T.; Bligny, R.; Maurel, C. Cytosolic pH regulates root water transport during anoxic stress through gating of aquaporins. *Nature* **2003**, *425*, 393−397.

(8) Neilson, J. E.; McCammon, J. A. Calculating pK_a values in enzyme active sites. *Protein Sci.* **2003**, *12* (9), 1894−1901.

(9) Manallack, D. T. The pK_a Distribution of Drugs: Application to Drug Discovery. *Perspect. Med. Chem.* **2007**, *1*, 25−38.

(10) Mitra, A.; Kesisoglou, F. Impaired Drug Absorption Due to High Stomach pH: A Review of Strategies for Mitigation of Such Effect To Enable Pharmaceutical Product Development. *Mol. Pharmaceutics* **2013**, *10* (11), 3970−3979.

(11) Li, X.; Cooper, H. A. Measurement of Drug Lipophilicity and pK_a Using Acoustics. *Anal. Chem.* **2012**, *84* (6), 2609−2613.

(12) Uddin, N.; Choi, T. H.; Choi, C. H. Direct Absolute pK_a Predictions and Proton Transfer Mechanisms of Small Molecules in Aqueous Solution by QM/MM-MD. *J. Phys. Chem. B* **2013**, *117*, 6269−6275.
(13) Jorgensen, W. L.; Briggs, J. M.; Gao, J. *A Priori* Calculations of pK_a's for Organic Compounds in Water. The pK_a of Ethane. *J. Am. Chem. Soc.* **1987**, *109*, 6857−6858.
(14) Briggs, J. M.; Jorgensen, W. L. *A Priori* pK_a Calculations and the Hydration of Organic Anions. *J. Am. Chem. Soc.* **1989**, *111*, 4190−4197.
(15) Lim, C.; Bashford, D.; Karplus, M. Absolute pK_a calculations with continuum dielectric methods. *J. Phys. Chem.* **1991**, *95*, 5610−5620.
(16) Ho, J.; Coote, M. L. A universal approach for continuum solvent pK_a calculations: are we there yet? *Theor. Chem. Acc.* **2010**, *125*, 3−21.
(17) Kallies, B.; Mitzner, R. pK_a Values of Amines in Water from Quantum Mechanical Calculations Using a Polarized Dielectric Continuum Representation of the Solvent. *J. Phys. Chem. B* **1997**, *101*, 2959−2967.
(18) Liptak, M. D.; Shields, G. C. Accurate pK_a Calculations for Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with CPCM Continuum Solvation Methods. *J. Am. Chem. Soc.* **2001**, *123*, 7314−7319.
(19) Dong, H.; Du, H.; Qian, X. Prediction of pK_a Values for Oligomethacrylic Acids Using Combined Classical and Quantum Approaches. *J. Phys. Chem. B* **2009**, *113*, 12857−12859.
(20) Warwicker, J. pK_a predictions with a coupled finite difference Poisson−Boltzmann and Debye−Hückel method. *Proteins: Struct. Funct. Bioinf.* **2011**, *79* (12), 3374−3380.
(21) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. M. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.
(22) Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graphics* **1990**, *8* (1), 52−56.
(23) Burgi, R.; Kollman, P. A.; van Gunsteren, W. F. Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation. *Proteins: Struct. Funct. Gen.* **2002**, *47* (4), 469−480.
(24) Aleksandrov, A.; Polydorides, S.; Archontis, G.; Simonson, T. Predicting the Acid/Base Behavior of Proteins: A constant-pH Monte Carlo Approach with Generalized Born Solvent. *J. Phys. Chem. B* **2010**, *114* (32), 10634−10648.
(25) Kong, X.; Brooks, C. L. λ-dynamics: A new approach to free energy calculations. *J. Chem. Phys.* **1996**, *105* (6), 2414.
(26) Guo, Z.; Brooks, C. L.; Kong, X. Efficient and flexible algorithm for free energy calculations using the λ-dynamics approach. *J. Phys. Chem. B* **1998**, *102*, 2032−2036.
(27) Knight, J. L.; Brooks, C. L. λ-dynamics free energy simulation methods. *J. Comput. Chem.* **2009**, *30*, 1692−1700.
(28) Swails, J. M.; Roitberg, A. E. Enhancing Conformation and Protonation State Sampling of Hen Egg White Lysozyme Using pH Replica Exchange Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8* (11), 4393−4404.
(29) Morrow, B. H.; Koenig, P. H.; Shen, J. K. Atomistic simulations of pH-dependent self-assembly of micelle and bilayer from fatty acids. *J. Chem. Phys.* **2012**, *137* (19), 194902.
(30) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys.* **2002**, *117* (9), 4184.
(31) Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. M. Coupling Constant pH Molecular Dynamics with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2010**, *6*, 560−568.
(32) Williams, S. L.; Blachly, P. G.; McCammon, J. A. Measuring the successes and deficiencies of constant pH molecular dynamics: A blind prediction study. *Proteins: Struct. Funct. Bioinf.* **2011**, *79* (12), 3381−3388.
(33) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **2004**, *25* (16), 2038−2048.
(34) Mongan, J.; Case, D. A. Biomolecular simulations at constant pH. *Curr. Opin. Struct. Biol.* **2005**, *15* (2), 157−163.
(35) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(36) Kong, X.; Brooks, C. L. λ-dynamics: a new approach to free energy calculations. *J. Chem. Phys.* **1996**, *105*, 2414−2423.
(37) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins: Struct. Funct. Bioinf.* **2004**, *56* (4), 738−752.
(38) Simonson, T.; Carlsson, J.; Case, D. A. Proton Binding to Proteins: pK_a Calculations with Explicit and Implicit Solvent Models. *J. Am. Chem. Soc.* **2004**, *126* (13), 4167−4180.
(39) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306−317.
(40) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26* (16), 1701−1718.
(41) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(42) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656−1676.
(43) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21* (12), 1049−1074.
(44) Sorin, E. J.; Pande, V. S. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophys. J.* **2005**, *88* (4), 2472−2493.
(45) DePaul, A. J.; Thompson, E. J.; Patel, S. S.; Haldeman, K.; Sorin, E. J. Equilibrium conformational dynamics in an RNA tetraloop from massively parallel molecular dynamics. *Nucleic Acids Res.* **2010**, *38* (14), 4856−4867.
(46) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(47) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81* (8), 3684−3690.
(48) Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(49) Migamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13* (8), 952−962.
(50) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2007**, *4*, 116−122.
(51) Orufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct. Funct. Bioinf.* **2004**, *55*, 383−394.
(52) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(53) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(54) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.
(55) Berendsen, H. J. C. *Simulating the Physical World*; Cambridge University Press: Groningen, Netherlands, 2007; pp 223−224.
(56) Case, D. A. *AMBER 11*; University of California, San Francisco.
(57) pK_a compilation. research.chem.psu.edu/brpgroup/pka_compilation.pdf (accessed Feb 17, 2014).

(58) Pearlman, D. A. Evaluating the Molecular Mechanics Poisson–Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase. *J. Med. Chem.* **2005**, *48*, 7796–7807.
(59) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2011**, *51* (1), 69–82.
(60) Balogh, G. T.; Gyarmati, B.; Nagy, B.; Molnar, L.; Keseru, G. M. Comparative Evaluation of *in Silico* pK_a Prediction Tools on the Gold Standard Dataset. *QSAR Comb. Sci.* **2009**, *28* (10), 1148–1155.
(61) Bortalato, A.; Tehan, B. G.; Bodnarchuk, M. S.; Essex, J. W.; Mason, J. S. Water Network Perturbation in Ligand Binding: Adenosine A2A Antagonists as a Case Study. *J. Chem. Inf. Model.* **2013**, *53* (7), 1700–1713.
(62) Bodnarchuk, M. S.; Viner, R.; Michel, J.; Essex, J. W. Strategies to calculate water binding free energies in protein-ligand complexes. *J. Chem. Inf. Model.* **2014**, DOI: 10.1021/ci400674k.
(63) Woods, C. J.; King, M. A.; Essex, J. W. The Development of Replica-Exchange-Based Free-Energy Methods. *J. Phys. Chem. B* **2003**, *107* (49), 13703–13710.
(64) Woods, C. J.; King, M. A.; Essex, J. W. Enhanced Configurational Sampling in Binding Free-Energy Calculations. *J. Phys. Chem. B* **2003**, *107* (49), 13711–13718.
(65) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607.
(66) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910.