

# pKa Calculations of Asp26 in Thioredoxin with Alchemical Free Energy Simulations and Hirshfeld-I Atomic Charges.

Aharon Gomez - Esteban Vöhringer-Martinez\*

October 21, 2018

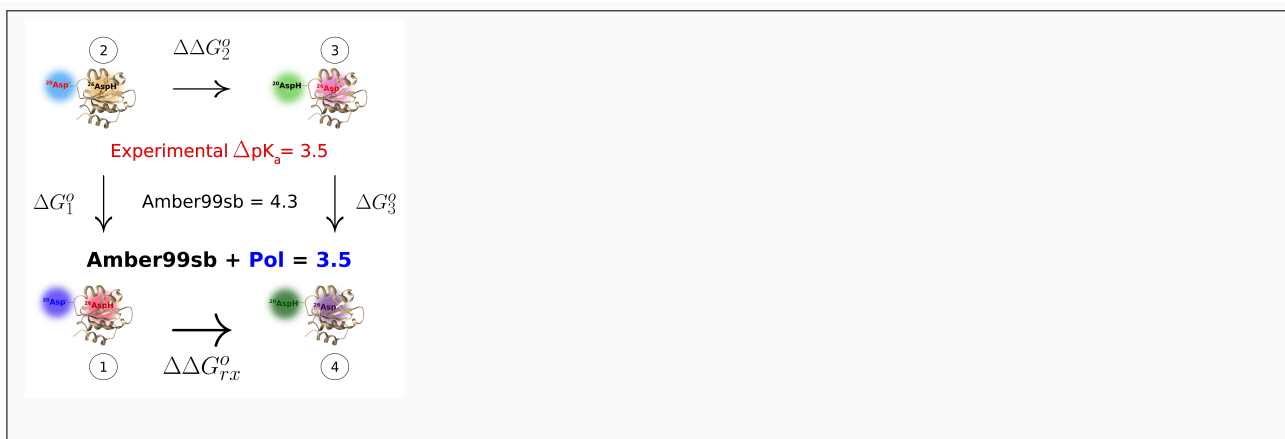
## Abstract

Thioredoxin is a protein that has been used as model system by various computational methods to predict the  $pK_a$  of aspartate residue Asp26 which is 3.5 units higher than the solvent exposed Asp20. Here, we use extensive atomistic molecular dynamics simulations of two different protonation states of Asp26 in combination with conformational analysis based on RMSD clustering and principle component analysis to identify representative conformations of the protein in solution. For each conformation the Gibbs free energy of proton transfer between the two aspartic acid residues is calculated with the Amber99sb force field in alchemical transformation. The varying polarization of Asp20/26 in different molecular environments and protonation states is described by Hirshfeld-I (HI) atomic charges obtained from the averaged polarized electron density. Our results show that the Gibbs free energy of proton transfer is dependent on the protein conformation, the proper sampling of the neighbouring Lys57 positions and on water molecules entering the hydrophobic cavity upon deprotonating Asp26. The inclusion of polarization of both aspartate residues in the free energy cycle by the HI atomic charges improve the results from the nonpolarizable force field and reproduces the experimental reference  $\Delta pK_a$  value.

**Keywords:** Free energy calculations, Hirshfeld-I atomic charges, MBIS atomic charges, pKa, polarization. ■

---

\*Departamento de Físico-Química, Facultad de Ciencias Químicas, Universidad de Concepción, Chile, Tel.: +56-41-2204986, evohringer@udec.cl



# INTRODUCTION

The acidity constant  $K_a$  of ionizable residues in proteins is a physico-chemical property of great relevance to biological systems. Changes in the protonation states of acidic or basic residues alter the active conformations of proteins<sup>1</sup>, their stability and solubility<sup>2</sup>, and is directly related to the catalytic efficiency of enzymes<sup>3</sup>. Prediction of the acidity constant – or its negative logarithm  $pK_a$  – for ionizable side chains in proteins, therefore, is currently a great challenge and has been addressed in the past with a variety of computational methods<sup>4</sup>.

Computational methods used to calculate the  $pK_a$  are based on the Gibbs free energy change between the protonated and deprotonated state of a residue in a protein. Most common methods are empirical to achieve computational efficiency; e.g. the software PROPKA.<sup>5</sup> They use continuum electrostatics and additional terms accounting for hydrogen bonds, desolvation free energies and other types of interactions. These methods are very useful for protein residues exposed to solvents but fail in cases when residues are buried in hydrophobic cavities.<sup>6</sup> To achieve their computational speed only one conformation (X-Ray or NMR structures) is considered without including changes in the conformations for different protonation states. To correct for local conformational changes the MCCE method has been proposed that is also based on continuum electrostatics and accounts for side chain rotamers in a Monte Carlo type approach and obtains the  $pK_a$  from titration curves.<sup>7</sup>

All continuum electrostatics methods neglect the contribution of water molecules to the Gibbs free energy change, especially the solvent entropy. Alchemical free energy calculations with an atomistic description of the protein, the surrounding water and ions are more rigorous. In these alchemical transformations in which a net charge is generated or annihilated a correct description for the long range electrostatics in the molecular dynamics simulation is required to avoid artifacts arising from systems with a net charge in combination with the Ewald summation.<sup>8</sup> When a complete sampling of the configurational space in the atomistic molecular dynamics simulations and reversibility in the alchemical transformations is achieved the accuracy of these methods is only limited by the force fields used to describe the interactions between the atoms.

For ionizable residues bearing a net negative or positive charge in their deprotonated or

protonated state electrostatic interactions are dominant and in non-polarizable force fields these type of interactions are described by atomic charges. In some force fields these atomic charges are optimized to reproduce hydration free energies (OPLS)<sup>9</sup> or interaction energies with some discrete water molecules for small representative molecules (CHARMM)<sup>10</sup>. The AMBER force field<sup>11</sup> has adopted a different approach and derives atomic charges with the RESP method that employs the molecular electrostatic potential obtained from HF/6-31G(d) electronic structure calculation in vacuum.<sup>12</sup> All these methods, however, do not include the varying polarization in different molecular environments. Especially in heterogeneous systems as proteins varying polarization of solvent exposed residues or the ones in hydrophobic cavities should be included in the electrostatic interactions, yet maintaining the simple non-polarizable character of traditional force fields for computational efficiency. An alternative approach is to describe the polarizability in the force field by atom induced dipole moments as in the AMOEBA force field<sup>13</sup> or the Drude oscillator method in the CHARMM polarizable force field<sup>14</sup>. This inclusion of polarizability, however, comes with a computational cost of minimizing the mutual polarization of the involved atomic induced dipole moments or the Drude particles.

One alternative approach that has been recently proposed uses the protein specific charges (PPC) which are calculated with a fragmented-based quantum chemistry approach and an implicit continuum solvation model<sup>15</sup>. As in the AMBER force field<sup>11</sup> atomic charges are also derived with the RESP method but the atomic charges of the rest of the protein are included as surface charges in the continuum solvation model that polarizes the electron density. This polarized electron density provides the molecular electrostatic potential used to optimize the charges. One problem of these atomic charges is that they are calculated for one conformation only and neglect specific hydrogen bonds with water molecules.

Here, we propose to use our recently introduced dynamic Hirshfeld-I charges<sup>16–21</sup> to describe the electrostatic interactions of ionizable residues in proteins. Hirshfeld-I charges are based on a stockholder partitioning of the molecular electron density by atomic densities of pro-atoms, charges of which are consistent with the charge of the atom in the molecule. These atomic charges have been shown to reproduce the molecular electrostatic potential of several organic molecules in the gas phase and present only a minor dependence on the molecular

conformation and the electronic structure method<sup>22</sup>. Recently, we have shown that they are also suitable for condensed phases and molecular dynamics simulations reproducing partition coefficients of DNA bases and hydration coefficient of 613 organic molecules in the FreeSolv database<sup>19,21</sup>. Because the molecular environment alters the molecular electron density of ionizable residues and these variations may depend on the protein conformation we employ our recently proposed dynamic method<sup>16-18,20</sup>, which describes the molecular environment provided by the protein and the solvent explicitly in QM/MM calculation. For the rest of the atomic interactions (Lennard-Jones potential parameters and the bonded interactions) the parameters in the AMBER99sb force field has been used to maintain compatibility.<sup>23</sup>

The performance of dynamic Hirshfeld-I charges to describe electrostatic interaction in ionizable residues was addressed by calculating the change in  $pK_a$  of the aspartic acid residue Asp26 in Thioredoxin.<sup>24</sup> This residue has a  $pK_a$  3.5 units higher than the same residue (e.g. Asp20) exposed to solvent ( $pK_a=4.0$ ).<sup>25-27</sup> Several computational methods have been applied to address the  $pK_a$  change of this residue in the last 15 years<sup>15,28,29</sup>. Simonson *et. al.* studied the system using non polarizable force fields in combination with implicit and explicit solvent and reported that the  $pK_a$  depends on the force field used in the simulation and on the description of the solvent by an implicit or explicit model.<sup>28</sup> For the explicit description of the solvent they conclude that including polarization effects in modeling Asp26 would improve the description of the salt bridge formed by Asp26 and lysine 57 (Lys57). They also pointed out that conformational sampling of both protonation states are required to achieve a proper relaxation of the solvent molecules. Ji and co-workers account for the polarization of Asp26 by the above mentioned PPC method and followed the same protocol as Simonson for the free energy calculations.<sup>15</sup> Their results were closer to the experimental reference, however, in both studies a net charge in the simulation box was generated which combined with the Ewald method to treat the long range electrostatics produces an artifact in the free energy.<sup>8</sup> Burger and co-workers used a QM/MM method to include the polarization of the atomic charges of Asp26 with the RESP method and studied the variation of  $\Delta pK_a$  with the salt bridge formation to Lys57 using implicit solvent.<sup>29</sup> They report that the protonation state of Asp26 alters protein conformation and this may also influence the free energy associated with its  $pK_a$ .

The importance of conformational changes in different protonation states and their effect on  $\Delta pK_a$  calculations has been recently addressed in staphylococcal nuclease mutants by Zheng and Cui with microsecond molecular dynamics simulation and polarizable and non-polarizable force fields.<sup>30</sup> The authors report that changes in the protonation state of buried titratable residues induce loss of secondary structure, rotations of side chains to be more solvent exposed and varying number of water molecules in hydrophobic pockets, which is only observed at relatively long time scales. Including this conformational changes the authors could improve the  $pK_a$  prediction with Poisson-Boltzmann calculations. In a subsequent study by Liu *et.al.* pH replica exchange molecular dynamics simulations with an implicit solvation model (GB) were used to obtain titration curves of the buried residues in the same staphylococcal nuclease mutants.<sup>31</sup> The study also concludes that varying protonation states induce conformational changes of the protein and that accounting for these conformations is crucial to achieve a good match with experimental  $pK_a$  values.

Based on the importance of protein conformation observed in staphylococcal nuclease mutants, in this study we use microsecond molecular dynamics simulations to assess the different configurations adopted by Thioredoxin for different protonation states of Asp26 employing the AMBER99sb force field. The most representative structures based on RMSD cluster and principal component analysis were then used to calculate dynamic Hirshfeld-I atomic charges that account for the averaged polarization of the electron density of Asp26 and Asp20 that is solvent exposed. These atomic charges improve the description of the electrostatic interactions of these charged residues in the protonated and also deprotonated form. Finally, we use alchemical free energy calculations to obtain the  $\Delta\Delta G^\circ$  value to transfer the proton from Asp26 to Asp20 and therefore the  $\Delta pK_a$  value between the two residues, which should match the experimental value of 3.5 (this approach maintains the simulated system neutral during the alchemical transformation and avoids the problem described above).

## METHODOLOGY

The  $\Delta pK_a$  value was calculated from proton transfer free energy between Asp26 and Asp20 for each representative conformation obtained from the cluster analysis method. These

free energies are calculated alchemically at the force field level and then the free energy to polarize the residue and adopt the atomic charges obtained from the residue's molecular electron density polarized by its environment is included in the free energy cycle as shown in Figure 1. The following subsections describe the setup of the simulations system, the method to obtain the polarized atomic charges and the details of the free energy calculations.

**System Setup** The starting point of this study was the crystalline structure 2TRX<sup>32</sup> obtained at pH 3.9 and 4°C. Two systems were considered denoted State A when Asp26 is protonated and Asp20 deprotonated and State B when both residues are deprotonated. For each of them, the protein was centered in a dodecahedron box solvated with  $\sim 4300$  water molecules described by the TIP3P<sup>33</sup> model and 4 or 5 Na ions to compensate the overall charge of the system. Cysteines were reduced and the only histidine single protonated. After energy minimization with the AMBER99sb force field, the whole system was equilibrated in the NVT and subsequently in the NPT ensemble for 100ps at 298 K and 1 bar. The simulations were performed with the Gromacs 5.1.4 software package<sup>34</sup> using a time step of 2fs in combination with molecular dynamics, velocity rescaling thermostat<sup>35</sup> and the Parrinello-Rahman<sup>36</sup> pressure coupling ( $\tau_p = 2ps$ ) algorithm using the compressibility of water. The electrostatic interactions were calculated with the Particle–Mesh–Ewald method, a cut-off radius of 1.0 nm, pme-order of 4 and a spacing of 0.16 nm.

Each state A and B differing in the protonation state of Asp26 (A protonated, B deprotonated) was simulated for 2  $\mu s$  and the sampling convergence of the conformational space was verified by projecting the trajectories at 0.5, 1 and 1.5  $\mu s$  on the two first eigenvectors of a principal component analysis of the backbone atoms RMSD of the whole trajectory. Additionally, a clustering method based on the backbone atom RMSD was performed with the GROMOS algorithm to identify the most representative conformations and their population<sup>37</sup>. The percentage of structures belonging to the five most representative clusters were used as their Boltzmann weight  $\omega_i$ . The obtained structures from the clustering method were projected on the first three eigenvectors to verify that they span large parts of the conformational space. These three eigenvectors represent a large part of the protein motion as shown in the distribution of eigenvalues for each vector in Figure S1 of the Supporting

Information.

**Atomic Charges** For each representative conformation from the cluster analysis atomic charges of the side chain atoms of both aspartic acid residues (Asp26 and Asp20) were replaced by dynamic Hirshfeld-I atomic charges using an explicit atomistic description of the solvent and the surrounding protein as described in our previous studies<sup>16-21</sup>: the QM/MM methodology is used to calculate the electronic density of the residue polarized by the instantaneous configuration of the surrounding atoms and the calculated charges by the Hirshfeld-I method replace the old charges of the respective QM atoms in the force field for the subsequent steps of the molecular dynamics simulations. All residue atoms up to the  $\alpha$  carbon atom were included in the QM region adding hydrogen link atoms to saturate its valence at the equilibrium bond distance. The rest of the force field parameters of residue Asp26 or Asp20 were taken from the AMBER99sb force field. The atomic charge of each atom is obtained as an average from the whole trajectory, varying its length (500ps-3000ps) to check upon convergence. The QM/MM molecular dynamics simulations were performed in combination with the ORCA package 3.0.2<sup>38</sup> in an electrostatic embedding as incorporated in the Gromacs 4.5.6<sup>39</sup> package with the simulation parameters described below. The obtained charges were also averaged on atoms which had the same value in the original AMBER99sb due to symmetric considerations or free rotations and the charge on the  $\beta$  carbon atom was adjusted to match the overall charge of the residue in the AMBER99sb force field. Atomic charges of the rest of the protein atoms in the AMBER99sb force field were maintained because they correctly describe the secondary structure of proteins.

**Free Energy Simulations and  $\Delta pK_a$**  The  $\Delta pK_a$  was obtained from the respective Gibbs free energy of proton transfer of each conformation from the cluster analysis in state A employing state-of-the art free energy calculations, minimizing sampling and free energy estimate errors. Hydrogen dummy atoms were added in the Asp20 and Asp26 residues, depending on the initial state, and energy minimized to use as the initial structure for the proton transfer free energy.

The Gibbs free energy of proton transfer was calculated starting from the equilibrated



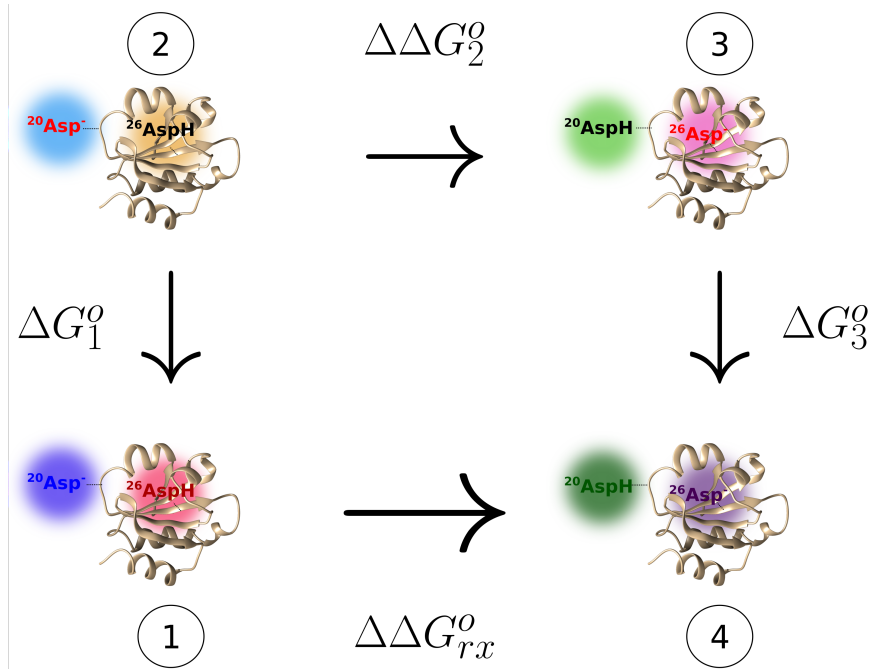


Figure 1: thermodynamic cycle

representative conformations described above in alchemical free energy calculations between states 2 and 3 (see Figure 1) with the AMBER99sb force field where the electrostatic interactions are switched between the protonated and deprotonated state for each aspartate residue using following  $\lambda$  parameters [1.00, 0.88, 0.75, 0.63 0.5, 0.38, 0.25, 0.13, 0.00]. The number of  $\lambda$  values and the simulation time per each value was optimized to reduce the Gibbs free energy estimation error. The analysis of the free energy simulations was performed with the Bennett–Acceptance–Ratio<sup>40</sup> method implemented in the Gromacs simulation package. These calculations provide the  $\Delta\Delta G_2^0$  values in Figure 1 since the atomic charges of the deprotonated or protonated form of Asp20 or Asp26 in the employed AMBER99sb force field do not change with the environment as is schematically represented with the same color used for the respective label of each residue in Figure 1.

States 2 and 3 were subsequently transformed to state 1 or 4 respectively by changing the Amber99sb atomic charges of both Asp20 and Asp26 residues to the dynamic Hirshfeld-I atomic charges which are polarized by the environment. These charges differ for the same protonation state of both residues depending on the environment, which is represented by different colors in their residue label in Figure 1. This change in free energy was denoted

$\Delta G_{1/3,\text{elec}}^0$  because it describes the variation in the electrostatic interactions due to polarization.

To obtain  $\Delta G_{1/3}^0$  in the free energy cycle the energy required to polarize the electron density has to be taken into account which is described in the following paragraph.

**Polarization correction to the free energy of proton transfer** To describe correctly the molecular environment of a residue it is necessary to take into account the polarization of the residue's electron density. If atomic charges are calculated in vacuum, this polarization contribution is neglected, assuming that the electronic density of the solute in the gas phase and in solution is the same. In this study, atomic charges are derived including the polarization from an explicit description of the solvent and the rest of the protein. Therefore, in the free energy cycle an additional energetic contribution  $\langle E_{\text{pol}} \rangle$  has to be added to  $\Delta G_{1/3,\text{elec}}^0$  which accounts for the ensemble averaged energy required to polarize both aspartic acid residues. The sum of  $E_{\text{pol}}$  and  $\Delta G_{1/3,\text{elec}}^0$  yields the value of  $\Delta G_{1/3}^0$  shown in Figure 1.

To calculate the energetic cost associated with the residue polarization  $E_{\text{pol}}$ , the electronic structure Hamiltonian of the vacuum calculation  $\hat{H}_{\text{vac}}$  was applied on the wave function of each solute that was polarized by the reaction field of the solvent described by the QM/MM method. The energetic cost is the difference between the expectation value of this calculation and the self-consistent-field energy obtained in the calculation in vacuum<sup>41</sup>.

$$E_{\text{pol}} = \left\langle \Psi_{\text{pol}} | \hat{H}_{\text{vac}} | \Psi_{\text{pol}} \right\rangle - \left\langle \Psi_{\text{vac}} | \hat{H}_{\text{vac}} | \Psi_{\text{vac}} \right\rangle \quad (1)$$

The energy to polarize the electron density  $\langle E_{\text{pol}} \rangle$  was obtained as an ensemble average of the  $E_{\text{pol}}$  values from the trajectory used to derive the charges, where the  $\Psi_{\text{pol}}$  wave function in the equation above was obtained from each QM/MM calculation, including the electrostatic interaction with the molecular environment. To avoid overpolarization of the hydrogen atoms used to saturate the QM system, the charge from the bonded backbone atoms was set to zero.

It is important to note that the derived atomic charges and the corresponding polarization correction to the proton transfer free energy are different for each aspartic acid in its molecular environment; a hydrophobic cavity or a solvent exposed residue.

Finally,  $\Delta\Delta G_{rx,i}^o$  was calculated for each conformation summing up  $\Delta G_1^0$ ,  $\Delta\Delta G_2^0$  and  $\Delta G_3^0$  to obtain  $\Delta\Delta G_{rx}^o$  according to Hritz<sup>42</sup> *et. al.* using the weight  $\omega_i$  from the cluster analysis for each conformation as

$$\Delta\Delta G_{rx}^o = -RT \ln \sum_{n=1}^5 \omega_i \exp \frac{-\Delta\Delta G_{rx,i}^o}{RT} \quad (2)$$

The  $\Delta pK_a$  is obtained as:

$$\Delta pK_a = \frac{\Delta\Delta G_{rx}^o}{2.303RT} \quad (3)$$

where  $R$  is the ideal gas constant and  $T$  the temperature in Kelvin (298 K).

Alternatively for the most representative conformation the sampling time per  $\lambda$  value was increased to 100ns assuring sampling of the slow rotation of the Lys57 residue during the transformation. In this case  $\Delta\Delta G_{rx}^o$  was calculated with one conformer as

$$\Delta\Delta G_{rx}^o = \Delta\Delta G_{rx,1}^o + RT \ln \omega_1 \quad (4)$$

where  $\Delta\Delta G_{rx,1}^o$  corresponds to the Gibbs transfer free energy with extended sampling of the most representative conformer of state A and  $\omega_1$  its weight from the cluster analysis.

## RESULTS

To address the change of  $pK_a$  between Asp26 and Asp20 in Thioredoxin first extensive molecular dynamics simulations of the protein were performed with Asp26 either protonated (state A) or deprotonated (state B). After conformational principle component analysis and clustering based on the backbone RMSD the representative conformations were used as starting point to calculate the free energy change associated with the proton transfer between Asp26 and Asp20 with alchemical free energy calculations and the AMBER99sb force field. The simulation time per alchemical state was optimized to sample also local slow conformational changes. Finally, the obtained free energies were corrected accounting for the polarization of both residues by the dynamic Hirshfeld-I atomic charges and the energy required to polarize the residue as shown in the free energy cycle in Figure 1.

# Conformational Analysis of Thioredoxin

Calculations of the  $pK_a$  of protein residues are highly dependent on the conformation of the protein and the description of the solvent.<sup>7,30,31</sup> Starting from the X-Ray structure Thioredoxin was simulated in explicit solvent for 2  $\mu$ s with Asp26 in the protonated (state A) or deprotonated form (state B). Slow conformational transitions related to changes in the position of backbone atoms were monitored by principal component analysis based on the backbone RMSD. The projection of the trajectory for different elapsed simulation times on the first two eigenvectors reveals that between 1.0 and 1.5  $\mu$ s the sampled phase space by the protein does not change and therefore most of the slow conformational transitions are well described in the total simulation time.

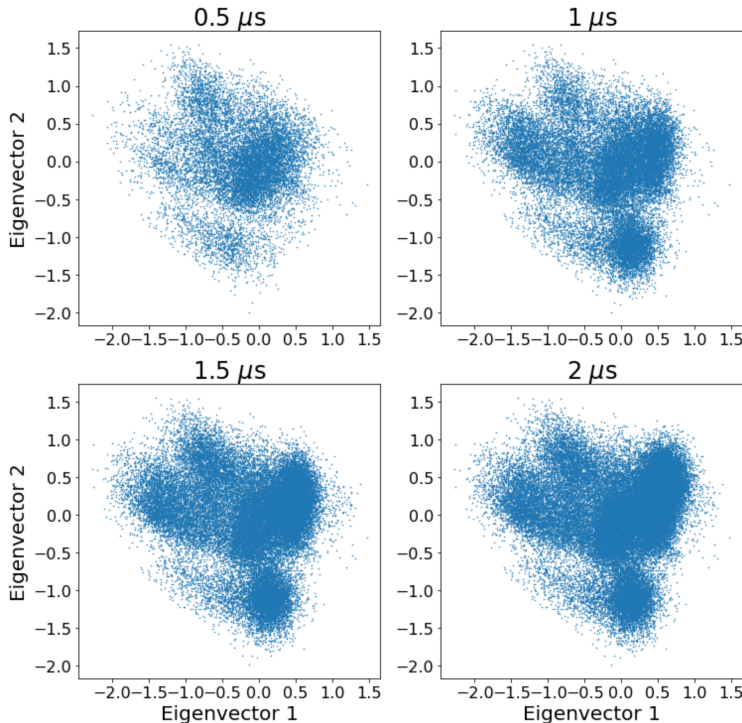


Figure 2: Projection of the trajectory of state A (Asp26 protonated) on the first two eigenvectors at different elapsed simulation times.

The same holds for state B where Asp26 was deprotonated, but only after 1.5  $\mu$ s convergence of the sampled phase space was achieved (see Figure S2 Supporting Information). The X-Ray structure was crystallized at pH 4.0 where Asp26 due to its high  $pK_a$  is ex-

pected to be protonated. Because the simulation of state B with an unprotonated Asp26 used the same crystal structure as initial configuration longer equilibration of the system was required accompanied with water molecules entering the hydrophobic cavity of Asp26 as will be discussed below.

After having verified convergence in the backbone conformational transitions, the RMSD of the backbone was used as metric in a cluster analysis to identify the most representative conformations. From the total clusters in state A the five most abundant were selected which unifies almost 80% of the structures and are shown superimposed in Figure 3A. The main differences in secondary structure are observed in the N-terminal region, the following  $\beta$ -sheet in sequence and the  $\alpha$  helix in the lower left of Figure 3A. The local molecular environment of Asp26 differs slightly in altering rotations of the hydroxyl group of the carboxy group of Asp26 and what is most evident the position of the positively charged Lys57 residue. This residue has been reported to play a key role in modulating the obtained  $pK_a$  by free energy calculations in implicit solvent<sup>29</sup>. The relative position of this residue with respect to Asp26 is most dependent on the dihedral angle  $\phi$  defined in part A of Figure 3. In part B of the same figure the value of this angle is shown as function of simulation time and many transition are observed adopting  $\phi$  values between 50 200 and 310°. The most abundant value lies at 50° and coincides with the value in the representative conformation of the cluster analysis possessing the largest population (64%) of structures shown in yellow in Figure 3A.

The obtained representative conformations were projected in combinations of the three largest eigenvectors from the PCA as shown in Figure 4. The most abundant conformation (yellow square) is located in the most dense and largest region of the phase space representation over the three projections. PCA and cluster analysis therefore confirm that this conformation is the most representative. The other four less abundant representative conformations are distributed in other regions of the sampled phase space with an almost complete coverage. The third most abundant conformation shown in violet is located in a region that is different from the main sampled region. Finally, the less abundant conformation (red) does not belong to a dense part of phase space which confirms its small Boltzmann weight obtained in the cluster analysis. Interestingly, the X-Ray structure shown as black star is not close to any abundant conformation identified in the cluster analysis, which may explain

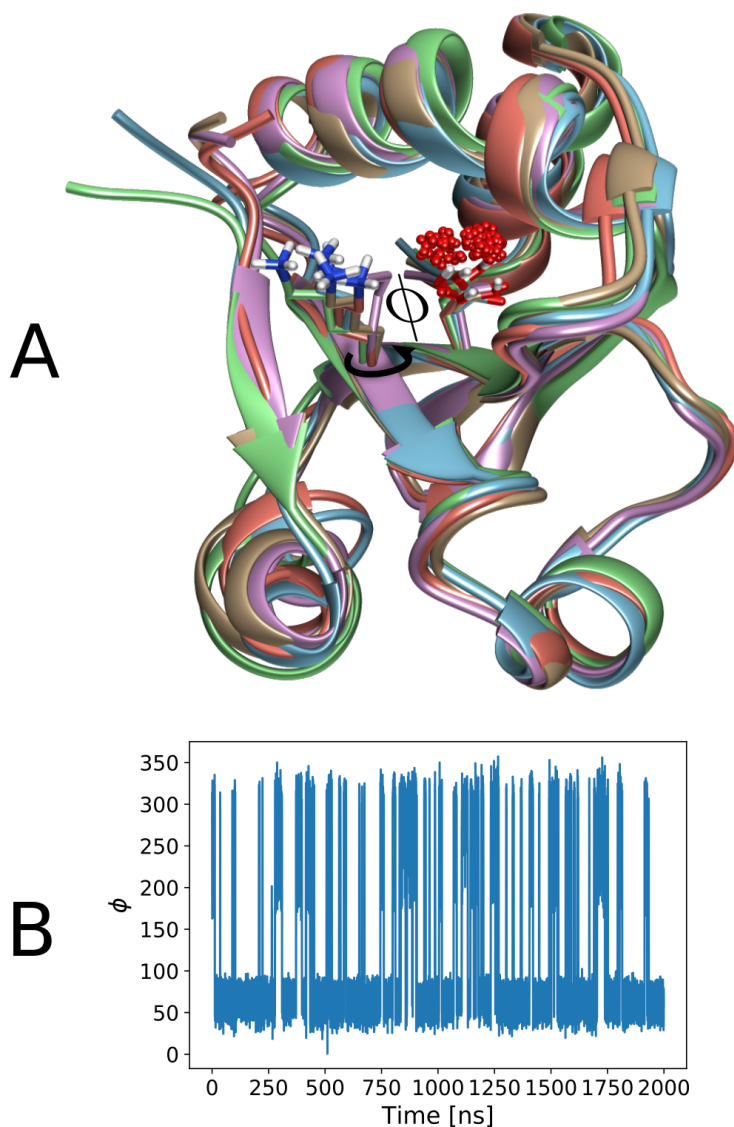


Figure 3: Figure A: Representative conformations of Thioredoxin in state A (Asp26 protonated) obtained from RMSD based cluster analysis color coded according to decreasing Boltzmann weight: yellow(64%), blue(14%), violet (11%), green (7%) and red (4%). The red spheres denote the hydration sites within a distance of 5 Å of Asp26. Figure B:  $\Phi$  dihedral angle of Lys57 defined in part A as function of simulation time.

the observed deviations in previous calculated  $pK_a$  changes employing this structure. The same analysis is also performed for state B obtaining the same conclusions (see Figure S3 in the Supporting Information)

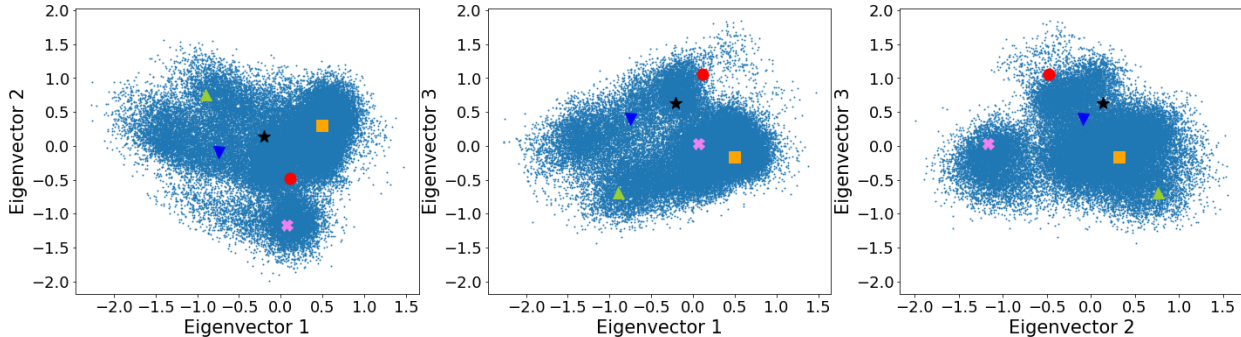


Figure 4: Projections of the representative conformations from the cluster analysis on combinations of the three largest principal eigenvectors. In order of decreasing abundance and in accordance to the colors used in Figure 3A: yellow square, blue triangle, violet cross, green triangle and red circle. The black star denote the crystal structure as reference.

## Gibbs Proton Transfer Free Energies

With the most representative conformations of state A (Asp26 protonated) the Gibbs free energy associated with the transfer of the proton of Asp26 to the solvent exposed Asp20 residue was calculated with alchemical free energy calculations using the standard AMBER99sb force field. In the standard force field the protonated or deprotonated form of the two aspartates are equivalent and independent of the molecular environment. The proton transfer free energy describes, therefore, the relative change in the interactions with the surrounding atoms of each protonation state maintaining the equivalency for the deprotonated or protonated form in different environments. In the free energy cycle this free energy corresponds to  $\Delta\Delta G_2^\circ$  in Figure 1 where the equivalency of Asp26 or Asp20 in each protonation states is represented as same colors.

The proton transfer free energy was calculated with several approaches to reduce its error due to incomplete sampling of relevant conformations on the alchemical path. Eight not

equally distributed  $\lambda$  values between zero and one reduced the free energy estimation error by the Bennett acceptance ratio (BAR) method considerably. To optimize the computational efficiency of the simulation protocol first we transferred the proton taking as starting structure for each  $\lambda$  value the same initial representative conformation described above, but the free energies presented an error of up to 6 kJ/mol (see Table S1 Supporting Information). To reduce the error the starting structure for each  $\lambda$  value was changed to the output structure of the  $\lambda - 1$  trajectory, which had a total simulation time of 10 ns. This method, which we name the quasi-slow growth method, provides a sequential local relaxation of the environment around the residues undergoing variation of their atomic charges and reduces the error in the free energy estimation to approximately 2 kJ/mol, which is acceptable to allow a good estimation of the  $pK_a$  (see Table S1 Supporting Information).

With the quasi slow growth method we calculated  $\Delta\Delta G_2^\circ$  (see Table 1) for each representative conformation of state A discussed above with 10 ns equilibration per  $\lambda$  window.  $\Delta\Delta G_2^\circ$  varies for each conformation confirming the relevance of an extensive conformational sampling of the protein for accurate  $\Delta pK_a$  predictions. In these simulations the slow degree of freedom represented by the angle  $\phi$  identified in the conformational analysis remained almost constant and the RMSD to the initial structure of state A for each conformation was below 1 Å (see Figure S4 and Table S3 Supporting Information). This indicates that in the alchemical transformation the protein remained in the same region of phase space. We then increased the simulation time to 100 ns per  $\lambda$  value for the two most abundant representative conformations (see Table 1). This allowed a proper sampling of the  $\phi$  angle during the alchemical transformation and reduced the value of  $\Delta\Delta G_2^\circ$ . The reduction is explained through a more stable deprotonated state which is only achieved through correct sampling of the  $\phi$  angle and probably allowing water molecules to enter the hydrophobic cavity.

If the transformation was performed in a reversible manner the conformations obtained after the proton transfer should be representative of state B where Asp26 is deprotonated (we assume that these conformations are independent of the protonation state of Asp20). Projection of the output conformations after the proton transfer on the three eigenvectors of state B show that they are very close to the most representative conformations of this state (see Figure S5 Supporting Information). The most representative conformations of



Table 1: Boltzmann weight  $w$  for each conformation of the cluster analysis calculated from the population of each cluster and its calculated proton transfer free energy change  $\Delta\Delta G_2^\circ$  in kJ/mol using the quasi slow growth method with different simulation times per  $\lambda$  window  $t_\lambda$ .

Conformation	$w$ (%)	$\Delta\Delta G_2^\circ$ ( $t_\lambda=10\text{ns}$ )	$\Delta\Delta G_2^\circ$ ( $t_\lambda=100\text{ns}$ )
1	64	$42.5 \pm 1.8$	$31.6 \pm 2.5$
2	14	$44.9 \pm 5.1$	$26.9 \pm 4.3$
3	11	$29.9 \pm 2.9$	
4	7	$38.8 \pm 0.9$	
5	4	$40.7 \pm 2.8$	

State B are shown in Figure 5 color coded according to the abundance together with the hydration sites around Asp26 for the most abundant conformation. In the conformation with the largest Boltzmann weight (yellow) Lys57 residue does not interact directly with Asp26, which is completely surrounded by hydration sites. But, this conformation was only obtained after a considerable simulation time since the starting structure corresponded to the X-Ray structure with Asp26 protonated at pH 4.0. Therefore to reach State B water has to enter the catalytic site and a conformational change in the Lys57 side chain has to take place. This confirms the observation in the alchemical transformation when the simulation time per  $\lambda$  window of 100 ns was employed (Figure S6 Supporting Information).

## Dynamic Hirshfeld-I (HI) Atomic Charges

For the most representative conformations of State A and B dynamic Hirshfeld-I atomic charges were derived which account for the averaged polarization of Asp26 and Asp20 in its respective environment (see methods). These atomic charges are derived from the polarized electron density of each residue (QM/MM methodology) for each configuration of a 3 ns simulation averaged over all values of each conformation of state A and state B. The atomic charges did not vary significantly between the conformations of each state and a Boltzmann weighted average value over all conformations is shown in Table 2 for each protonation state compared to the values in the Amber99sb force field (FF). The atomic charges on the carbon

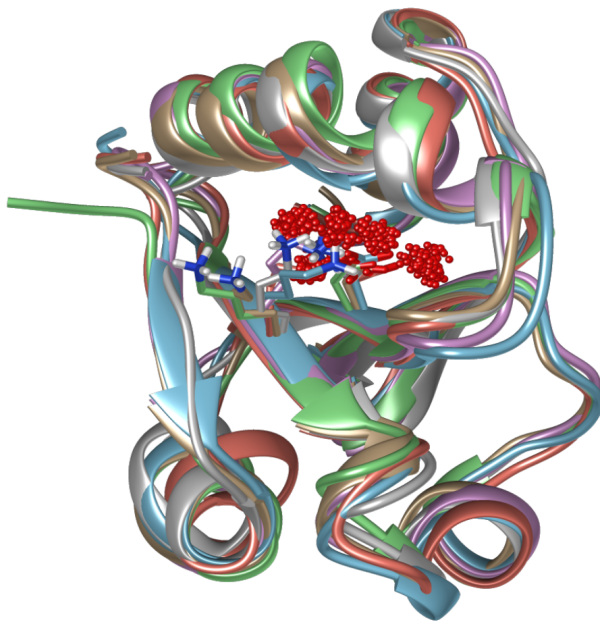


Figure 5: Representative conformations of Thioredoxin in state B (Asp26 deprotonated) obtained from RMSD based cluster analysis color coded according to decreasing Boltzmann weight: yellow(61%), blue(16%), violet (8%), green (6%), red (5%)) and white (4%). The red spheres denote the hydration sites within a distance of 5 Å of Asp26.

atoms are larger in absolute values than the one in Amber99sb because the RESP method employed in this force field adds an additional constraint on the carbon atom charge in the optimization to reproduce the molecular electrostatic potential. This results also in slightly larger values for the hydrogen atomic charges bonded to the  $\beta$  carbon atom (HB1 and HB2). For the protonated form of Asp26 (AspH26) in state A, where only some water molecules entered the hydrophobic cavity, only the atomic charge of the carbonyl oxygen atom (OD1) is slightly more negative than the one in the force field. The solvent exposed Asp20 residue, however, presents a significantly more negative charge on this atom and also the atomic charges of the hydroxyl group become larger in absolute value. Comparing these two residues the effect of the molecular environment becomes evident because Asp26 in a hydrophobic environment present atomic charges similar to the ones derived in vacuum used in the force field whereas the solvent exposed Asp20 is more polarized by the surrounding water molecules which is most evident in the carbonyl group forming specific residue solvent

Table 2: Dynamic Hirshfeld-I atomic charges on the side chain atoms of Asp26 and Asp20 in the protonated and deprotonated form. Uncertainty informed as standard deviation.

Atom type	FF	AspH(26)	AspH(20)	FF	Asp <sup>-</sup> (26)	Asp <sup>-</sup> (20)
CB	-0.0316	-0.4161 ± 0.03	-0.4861 ± 0.03	-0.0303	-0.2881 ± 0.03	-0.3281 ± 0.03
HB1	0.0488	0.1400 ± 0.01	0.1700 ± 0.02	-0.0122	0.1100 ± 0.01	0.1300 ± 0.01
HB2	0.0488	0.1400 ± 0.01	0.1700 ± 0.03	-0.0122	0.1100 ± 0.01	0.1300 ± 0.01
CG	0.6462	0.8900 ± 0.02	0.9400 ± 0.03	0.7994	0.9700 ± 0.03	0.9700 ± 0.03
OD1	-0.5554	-0.5900 ± 0.02	-0.6500 ± 0.03	-0.8014	-0.8800 ± 0.03	-0.8800 ± 0.03
OD2	-0.6376	-0.6300 ± 0.02	-0.6400 ± 0.03	-0.8014	-0.8800 ± 0.03	-0.8800 ± 0.03
HD2	0.4747	0.4600 ± 0.02	0.4900 ± 0.03			

Table 3: Energies in kJ mol<sup>-1</sup>

Contribution	AspH(26)	Asp <sup>-</sup> (26)	AspH(20)	Asp <sup>-</sup> (20)
$\Delta G_{elec}^o$	-91.2 ± 0.1	-187.2 ± 0.2	-157.9 ± 0.1	-220.4 ± 0.03
$\langle E_{pol} \rangle$	4.0 ± 0.04	38.7 ± 0.3	23.3 ± 0.3	33.9 ± 0.5
Total	-87.2 ± 0.14	-148.5 ± 0.5	-134.6 ± 0.4	-186.5 ± 0.53

hydrogen bonds. In the deprotonated state Asp26 and Asp20 present similar charges on the carboxy group which is associated with similar molecular environments. Indeed, as shown in the representative conformations of state B where Asp26 is deprotonated, the most representative conformation presents hydration sites which surround the carboxy group completely and provide similar type of interactions as Asp20 exposed to solvent. The obtained atomic charges are considerably larger than the force field ones because of the polarization by the aqueous environment.

The derived atomic charges have been used to evaluate the change in free energy  $\Delta G_1^0$  and  $\Delta G_3^0$  in the free energy cycle shown in Figure 1. These values are made up from the change in free energy associated with the alchemical transformation of the force field to the polarized dynamic Hirshfeld-I atomic charges of each residue  $\Delta G_{elec}^o$  which present a small dependence

on the protein conformation (see Table S2 Supporting Information) and the averaged energy required to polarize the electron density of the residue. The sum of these two contributions for each residue is shown in Table 3.  $\Delta G_{elec}^o$  values for the protonated form of Asp26 are smaller than Asp20 because the atomic charges of Asp26 in the hydrophobic cavity are smaller than the solvent exposed ones. Having atomic charges of Asp26 that are similar to the unpolarized ones in the force field due to the absence of the surrounding hydration sites also explains the small energy associated with the polarization of the electron density  $\langle E_{pol} \rangle$ . For the deprotonated Asp26 and Asp20 residue  $\Delta G_{elec}^o$  is more negative because both residues are exposed to the solvent increasing their absolute values of atomic charges which is accompanied with an increased polarization energy  $\langle E_{pol} \rangle$ . Asp20, however, presents a larger value for  $\Delta G_{elec}^o$  which is associated with a full solvent exposure whereas Asp26 reaches only a partial solvation as shown in Figure 5 in the most representative conformation.

## $\Delta pK_a$ Calculation

Table 4: Free energy contribution in  $\text{kJ mol}^{-1}$  in the thermodynamic cycle for each representative conformation. All  $\Delta\Delta G_{2\rightarrow 3,i}^o$  values were obtained with 10 ns sampling per  $\lambda$  value. Total free energy calculated as in reference<sup>42</sup>.

Conformation	$\chi_i(\%)$	$\Delta G_{1,i}^o$	$\Delta\Delta G_{2\rightarrow 3,i}^o$	$\Delta G_{3,i}^o$	$\Delta\Delta G_{rx,i}^o$
1	64	$-272.7 \pm 1.0$	$42.5 \pm 1.8$	$-283.0 \pm 0.8$	$32.2 \pm 3.6$
2	14	$-271.8 \pm 1.0$	$44.9 \pm 5.1$	$-283.5 \pm 0.7$	$33.2 \pm 6.8$
3	11	$-272.1 \pm 1.0$	$29.9 \pm 2.9$	$-282.7 \pm 0.8$	$19.3 \pm 4.7$
4	7	$-271.7 \pm 1.0$	$38.8 \pm 0.9$	$-282.5 \pm 0.9$	$28.0 \pm 2.8$
5	4	$-272.0 \pm 1.0$	$40.7 \pm 2.8$	$-283.6 \pm 0.8$	$29.1 \pm 4.6$
$\Delta\Delta G_{rx}^o = 24.6 \pm 4.5$					

The presented results of the free energy change  $\Delta G_{1,3}^0$  and  $\Delta\Delta G_2^o$  for each conformation (see Table 1) are combined in the cycle in Figure 1 to calculate the free energy change associated with the proton transfer from Asp26 to Asp20 accounting for the polarization of both residues in both protonation states and the different conformations. The associated

Table 5: Summary of  $\Delta pK_a$  values

	Standard Force Field		Polarization included	
	Single conformation	Multiple conformations	Single conformation	Multiple conformations
$\Delta pK_a$	$6.2 \pm 0.5$	$5.3 \pm 0.4$	$3.5 \pm 0.8$	$4.3 \pm 0.8$

change in free energy  $\Delta\Delta G_{rx}^o$  is calculated according to Hritz<sup>42</sup> *et. al.* using the weight of each conformation and its  $\Delta\Delta G_{rx,i}^o$  value as  $\Delta\Delta G_{rx}^o = -RT \ln \sum_{n=1}^5 \omega_i \exp \frac{-\Delta\Delta G_{rx,i}^o}{RT}$  as shown in Table 4. In this approach we assume that the phase space of the conformations of state A does not overlap with each other, which is confirmed by the small average RMSD value of the alchemical transformation to its respective initial structure (see Table S3). The obtained value shown in Table 4 deviates only by  $\simeq 4.6$  kJ/mol from the experimental reference value. Analyzing the values of  $\Delta G_{1,3}^0$  for the two end states reveals a small dependence on the conformation because the local environment in state A is maintained.

The above analysis is based on alchemical transformations without a complete sampling of the slow degrees of freedom ( $\phi$  angle etc.). Using 100 ns per  $\lambda$  window and the most abundant conformation  $\Delta\Delta G_2^o$  equals  $31.6 \pm 2.5$  kJ/mol which combined with the polarization cost of  $\Delta G_{1,3}^0$  results in  $\Delta\Delta G_{rx,1}^o = 21.3 \pm 4.3$  kJ/mol. Combining this value with the Boltzmann weight of this most representative conformation yields  $\Delta\Delta G_{rx}^o = \Delta\Delta G_{rx,1}^o + RT \ln \omega_1 = 20.2 \pm 4.3$  kJ/mol according to<sup>43</sup>, a value very close to the experimental reference.

Finally,  $\Delta pK_a$  is calculated with the obtained free energy values according to equation 3. As shown in Table 5 using the standard force field and the most abundant conformation sampling the slow degrees of freedom (100 ns per  $\lambda$ ) a deviation of 2.7  $pK_a$  units from the experimental value is observed, which is reduced by almost one unit when the free energy is calculated for each conformation independently in a multi conformational approach. When the polarization is taken into account in the free energy cycle the value for the most representative conformation matches the experimental reference and a deviation of 1 unit is observed in the multiconformational approach. Therefore, inclusion of an average polarization of the aspartate residues improved the  $\Delta pK_a$  considerably.

# CONCLUSIONS

The results show that the Gibbs free energy of proton transfer between Asp26 and Asp20 depends on the conformation of the protein, specifically on slow degrees of freedom as the  $\phi$  angle of Lys57 and on water molecules entering the hydrophobic cavity containing the deprotonated Asp26 during the alchemical transformation. The new proposed dynamic Hirshfeld-I atomic charges, which take an average polarization of the two residues into account, improve the calculated  $\Delta pK_a$  considerably achieving a perfect agreement with the experimental reference when the most representative conformation is used and the simulation time in the alchemical transformation is extended to account for slow degrees of freedom and water molecules entering the cavity.

# ACKNOWLEDGMENTS

The authors acknowledge financial support provided by Fondecyt No. 1160197.

## References

1. W.R. Forsyth, J.M. Antosiewicz, A.D. Robertson, *Proteins* 2002, **48**, 388.
2. C.N. Pace, G.R. Grimsley, J.M. Scholtz, *J. Biol. Chem.* 2009, **284**, 13285.
3. S. Long, X. Zhang, Z. Rao, K. Chen, M. Xu, T. Yang, S. Yang, *Enzyme Microb. Technol.* 2016, **82**, 15.
4. Z. Sun, X. Wang, J. Song, *J. Chem. Inf. Model.* 2017, **57**, 1621.
5. C.R. Sndergaard, M.H.M. Olsson, M. Rostkowski, J.H. Jensen, *J. Chem. Theory Comput.* 2011, **7**, 2284.
6. S. Witham, K. Talley, L. Wang, Z. Zhang, S. Sarkar, D. Gao, W. Yang, E. Alexov, *Proteins* 2011, **79**, 3389.
7. Y. Song, J. Mao, M.R. Gunner, *J. Comput. Chem.* 2009, **30**, 2231.
8. J.S. Hub, B.L. de Groot, H. Grubmiller, G. Groenhof, *J. Chem. Theory Comput.* 2014, **10**, 381.
9. W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, *J. Am. Chem. Soc.* 1996, **118**, 11225.
10. A.D. MacKerell Jr., D. Bashford, M. Bellott, R.L. Dunbrack Jr., J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wirkiewicz-Kuczera, D. Yin, M. Karplus, *J. Phys. Chem. B* 1998, **102**, 3586.
11. J. Wang, P. Cieplak, P.A. Kollman, *J. Comput. Chem.* 2000, **21**, 1049.
12. C.I. Bayly, P. Cieplak, W. Cornell, P.A. Kollman, *J. Phys. Chem.* 1993, **97**, 10269.
13. J.W. Ponder, C. Wu, P. Ren, V.S. Pande, J.D. Chodera, M.J. Schnieders, I. Haque, D.L. Mobley, D.S. Lambrecht, R.A. DiStasio, M. Head-Gordon, G.N.I. Clark, M.E. Johnson, T. Head-Gordon, *J. Phys. Chem. B* 2010, **114**, 2549.

14. A. Savelyev, A.D. MacKerell, *J. Comput. Chem.* 2014, **35**, 1219.
15. C. Ji, Y. Mei, J.Z.H. Zhang, *Biophys. J.* 2008, **95**, 1080.
16. D.A. Saez, E. Vöhringer-Martinez, *J. Comput. Aided Mol. Des.* 2015, **29**, 951.
17. E. Vöhringer-Martinez, C. Drner, *J. Phys. Chem. B* 2016, **120**, 12444.
18. E. Vöhringer-Martinez, T. Verstraelen, P.W. Ayers, *J. Phys. Chem. B* 2014, **118**, 9871.
19. M. Riquelme, A. Lara, D.L. Mobley, T. Verstraelen, A.R. Matamala, E. Vöhringer-Martinez, *J. Chem. Inf. Model.* 2018, **58**, 1779.
20. D.A. Saez, S. Vogt-Geisse, R. Inostroza-Rivera, T. Kuba, M. Elstner, A. Toro-Labb, E. Vöhringer-Martinez, *Phys. Chem. Chem. Phys.* 2016, **18**, 24033.
21. A. Lara, M. Riquelme, E. Vöhringer-Martinez, *J. Comput. Chem.* 2018, **39**, 1728.
22. S. Van Damme, P. Bultinck, S. Fias, *J. Chem. Theory Comput.* 2009, **5**, 334.
23. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins* 2006, **65**, 712.
24. J.F. Collet, J. Messens, *Antioxid. Redox Signal.* 2010, **13**, 1205.
25. M.A. Hass, F.A. Mulder, *Annu. Rev. Biophys.* 2015, **44**, 53.
26. K. Langsetmo, J.A. Fuchs, C. Woodward, *Biochemistry* 1991, **30**, 7603.
27. H.J. Dyson, M.F. Jeng, L.L. Tennant, I. Slaby, M. Lindell, D.S. Cui, S. Kuprin, A. Holmgren, *Biochemistry* 1997, **36**, 2622.
28. Thomas Simonson, J. Carlsson, , David A Case, *J. Am. Chem. Soc.* 2004, **126**, 4167.
29. S.K. Burger, J. Schofield, P.W. Ayers, *J. Phys. Chem. B* 2013, **117**, 14960.
30. Y. Zheng, Q. Cui, *Proteins* 2017, **85**, 268.
31. J. Liu, J. Swails, J.Z.H. Zhang, X. He, A.E. Roitberg, *J. Am. Chem. Soc.* 2018, **140**, 1639.



32. S.K. Katti, D.M. LeMaster, H. Eklund, *J. Mol. Biol.* 1990, **212**, 167.
33. W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, *J. Chem. Phys.* 1983, **79**, 926.
34. M.J. Abraham, T. Murtola, R. Schulz, S. Pll, J.C. Smith, B. Hess, E. Lindahl, *SoftwareX* 2015, **1-2**, 19.
35. G. Bussi, D. Donadio, M. Parrinello, *J. Chem. Phys.* 2007, **126**, 014101.
36. M. Parrinello, A. Rahman, *J. Appl. Phys.* 1981, **52**, 7182.
37. X. Daura, K. Gademann, B. Jaun, D. Seebach, W.F. van Gunsteren, A.E. Mark, *Angew. Chem.* 1999, **38**, 236.
38. F. Neese, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2012, **2**, 73.
39. B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, *J. Chem. Theory Comput.* 2008, **4**, 435.
40. C.H. Bennett, *J. Comput. Phys* 1976, **22**, 245.
41. A.R. da Cunha, E.L. Duarte, M.T. Lamy, K. Coutinho, *J. Chem. Phys* 2014, **440**, 69.
42. J. Hritz, C. Oostenbrink, *J. Phys. Chem. B* 2009, **113**, 12711.
43. S.C. Gill, N.M. Lim, P.B. Grinaway, A.S. Rustenburg, J. Fass, G.A. Ross, J.D. Chodera, D.L. Mobley, *J. Phys. Chem. B* 2018, **122**, 5579.