Consistent Approach for Calculating Protein pK_a's using Poisson-Boltzmann Model

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Han Wool Yoon

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Adviser: Yuk Sham

June 2013

Acknowledgements

First, I would like to thank Prof. Yuk Sham, my research advisor, for his guidance and support over the years. He has been an inspiration for me and I look forward to continue my journey in the field of computational biology at USC. I also would like to thank all the past and current members in the Prof. Sham’s lab who have been helpful in giving me valuable feedbacks in my research. I also wish to express my gratitude to my committee members, Dr. Elizabeth Amin and Dr. Carlos Sosa. I want to give special thanks to Melody Lee for editorial critiques of my thesis and her encouragement. Lastly, but not least, I want to thank my parents, Yong Chul Yoon and Eun Ja Park, and my fiancé, Michelle Sham, who have always supported me by all means throughout my academic career. My research could not have been completed without their support and patience.

Abstract

Accurate prediction of protein pK_a's is important to understand protein electrostatics and functions. Improving the accuracy of pK_a prediction using the Poisson-Boltzmann electrostatic model remains an active area of research. The major challenge is to determine the appropriate dielectric constant (ε_P) that best describes the heterogeneous protein environment. The common use of a single large ε_P often fails to reproduce large experimental pK_a shifts of biological important residues. In this study, I implemented a two steps approach, as described in earlier PDLD/S model, that uses a single low dielectric constant for calculating the intrinsic protein pK_a's when all other ionizable group are neutralized and a single large dielectric constant for evaluating the pK_a's shifts as a result of charge-charge coupling between ionizable groups. This approach is less sensitive to the dielectric constants used and can reliably reproduce the commonly observed protein pK_a's and others with abnormal large pKa shifts.

Table of Contents

Introductions

Method

Results and Discussions

List of Tables

List of Figures

Introductions

1.1 Ionizable residues in proteins

Protein electrostatics is an important factor governing the structural stabilities and functions of proteins.[1-4] A consistent model for predicting accurately the protein pK_a's can further provide the theoretical model for drug discovery and protein design applications. Proteins consist of amino acids with ionizable side chains that undergo proton association and dissociation reactions in aqueous solution. The pK_a's of these ionizable groups can be greatly influenced by their local environments such as the composition of the solvent mixture, pH and ionic concentration. In the unfolded state, the pK_a of these ionizable sidechains are presumed to be solvent exposed and their pK_a's are typically similar to that of the individual amino acids in aqueous solution. During the protein folding process, however, these ionizable groups become localized onto the surface or into the interior regions of the protein that engages in an intricate network of electrostatic and non-electrostatic interactions involving hydrogen bonding, charge-charge, charge-dipole, and hydrophobic interactions (Figure 1). These local heterogeneous environment greatly influences the overall energetics and stability of each ionizable residues and their corresponding apparent pK_a's in proteins. Ability to quantify the compensatory electrostatic and non-electrostatic effects provides a rigorous benchmark for examining protein electrostatics.

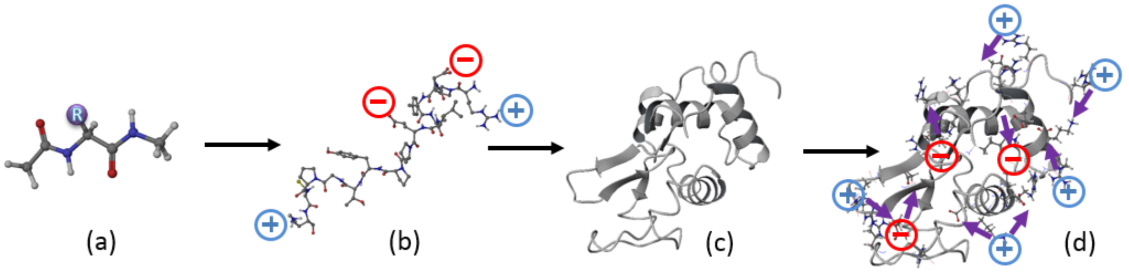


Figure 1-1 Amino acids(a) form peptide chains(b) followed by secondary and tertiary structures (c). Ionizable sites are localized into heterogeneous electrostatic environments (d).

1.2 Calculating pK_a in proteins

Understanding the role of electrostatic interactions in proteins is crucial for the study of biological functions. There are many essential biological processes that are modulated by the specific ionizable state of these ionizable groups. It governs the overall protein stability, folding pathway, ion transport, molecular association and catalysis. [1-3] Consistently and accurately quantifying the pK_a of ionizable groups and its specific ionization state in proteins is not a trivial task. Before the availability of 3 dimensional protein structures, early Tanford and Kirkwood (TK) model introduced for protein pKa calculation assumes proteins as an impenetrable macroscopic spheres consisting of a low dielectric hydrophobic protein core surrounded by ionizable group located on the surfaces of the protein.[5] In the era of X-ray protein crystallography that reveals many of these ionizable residues are buried, rigorous methods such as protein dipoles Langevin dipoles (PDLD) [6] , Poisson-Boltzmann (PB) [7-9] , and generalized Born (GB) [10, 11] type models have emerged that take into account the full atomistic detail of the protein structure, as well as the explicit and implicit representation of solvent environment. With increase in computational power, algorithm design and the number of high resolution crystallized protein structures available for examining protein pKa's, the accuracy in many of these computational models have significantly enhance over the past decades.

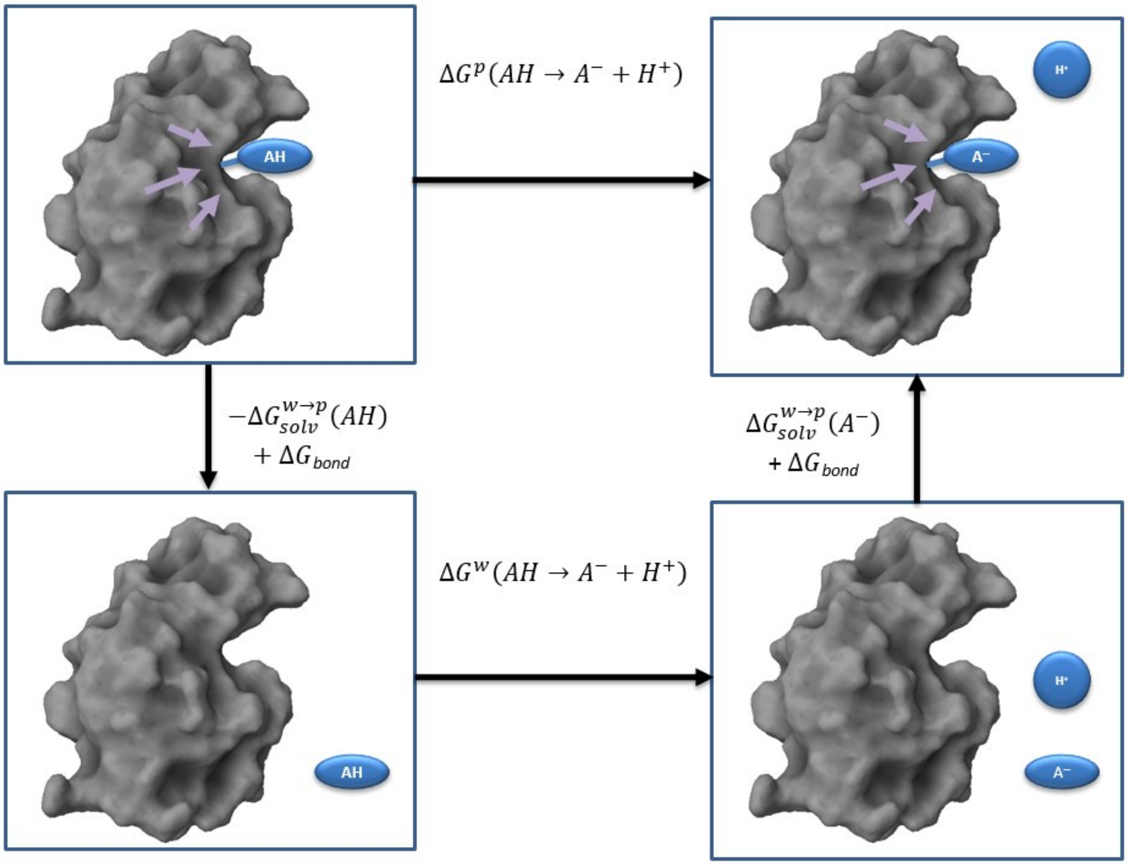

$\Delta G^{p}(AH \rightarrow A^{-} + H^{+})$
AH
A⁻
H⁺
$-\Delta G_{solv}^{w\rightarrow p}(AH) + \Delta G_{bond}$
$\Delta G_{solv}^{w\rightarrow p}(A^{-}) + \Delta G_{bond}$
$\Delta G^{w}(AH \rightarrow A^{-} + H^{+})$
AH
H⁺
A⁻

Figure 2. Thermodynamic cycle for predicting pK_a of an ionizable group in a protein. *w* and *p* designate water and protein, respectively. $\Delta G_{solv}^{w\rightarrow p}$ designates a change in solvation free energy of moving the titratable group from water to its protein.

The most widely used method for evaluating protein pKa is based on the thermodynamic cycle shown in Figure 2. Instead of directly calculating the change in free energy of deprotonation of the indicated ionizable group inside the protein, the method utilizes the deprotonation of the ionizable group in the aqueous phase as our reference reaction. This allows us to take the advantage of using the reference pK$_a$ values of ionizable amino acids which can be measured experimentally (Table 1). $\Delta G_{solv}^{w\rightarrow p}$ designates the change in solvation free energy of moving the titratable group from water (w) to its protein (p) environment. Both the ΔG_{bond} and the $\Delta G_{\mathrm{solv}}(\mathrm{H}^{+})$ terms are canceled from this cycle and the free energy difference of deprotonation of the side chain of the ionizable group can be given by

$$\Delta G^{p}(AH \rightarrow A^{-} + H^{+}) = \Delta G^{w}(AH \rightarrow A^{-} + H^{+}) + \Delta G_{solv}^{w\rightarrow p}(A^{-}) - \Delta G_{solv}^{w\rightarrow p}(AH) \quad (1.1)$$

which can be expressed in terms of pK$_a$ units as

$$\mathrm{pK_a^p(AH)} = \mathrm{pK_a^w(AH)} + \frac{1}{2.303\mathrm{RT}}\Delta\Delta\mathbf{G}_{\mathbf{solv}}^{\mathbf{w\rightarrow p}}(\mathbf{AH} \rightarrow \mathbf{A}^{-}) \quad (1.2)$$

Since we have the reference values for $\mathrm{pK_a^w(AH)}$, the only problem is to evaluate the change in the solvation energies of moving the protonated group from the protein to water and the deprotonated group from water to protein or vice versa depending on whether it is an acid or base. Both the PDLD and PB model which have been parameterized to reproduce the solvation free energy of small molecules and ions are described below.

Protein ionizable groups	$\boldsymbol{pK_a^{w,mod}}$
N-terminal NH3	*7.5 – 8.0*
C-terminal COO$^-$	*3.6 – 3.8*
Arginine	*12.0 – 12.5*
Aspartic Acid	*3.9 – 4.0*
Cysteine	*8.3*
Glutamic Acid	*4.3 – 4.4*
Histidine	*6.3 – 6.5*
Lysine	*10.4 – 10.5*
Tyrosine	*9.6*

Table 1 Model pK$_a$ of side chains of ionizable amino acids in water.

1.3 Poisson Boltzmann Electrostatic Model

Poisson-Boltzmann (PB) electrostatic continuum type models are one of the most popular methods for examining protein electrostatics. There are several implementations of PB model within popular software including Delphi [9, 12, 13], CHARMM [14], APBS[15], and Amber [16, 17] and web servers such as H++[18-20] and CHARMM-gui. [21] The Poisson-Boltzmann equation is given by

$$\nabla \cdot \varepsilon(\boldsymbol{r})\nabla\phi(\boldsymbol{r}) - \kappa^2\varepsilon(\boldsymbol{r})\sinh[\phi(\boldsymbol{r})] = -\boldsymbol{4}\pi\rho_0(\boldsymbol{r}) \qquad (1.3)$$

where $\phi(\boldsymbol{r})$ is the electrostatic potential that we need to calculate at distance *r*, $\rho_0(\boldsymbol{r})$ is the permanent charge density, $\varepsilon(\boldsymbol{r})$ is the distance dependent dielectric constant, and κ is the inverse Debye-Huckel salt screening length defined as

$$\kappa^2(\boldsymbol{r})\varepsilon(\boldsymbol{r}) = \frac{8\pi N_a e^2 I}{k_\beta T} \tag{1.4}$$

where N_a is the Avogadro's number, e is the electronic charge, and I is the ionic concentration. Using the Taylor series expansion, we can approximate $\sinh[\phi(\boldsymbol{r})]$ as $\phi(\boldsymbol{r})$ giving the linearized Poisson Boltzmann (LPB) equation as

$$\nabla \cdot \varepsilon(\boldsymbol{r})\nabla\phi(\boldsymbol{r}) - \kappa^2\varepsilon(\boldsymbol{r})\phi(\boldsymbol{r}) = -4\pi\rho_0(\boldsymbol{r}) \tag{1.5}$$

which can be calculated more rapidly. Since proteins are irregularly shaped, the PB equation can also be solved numerically with several discretization methods commonly referred as finite-difference Poisson-Boltzmann (FDPB) method. The implementation of Poisson-Boltzmann models is described in Figure 3. The space grid is built around the protein with each grid point represents by a polarizable implicit solvent molecule with a water dielectric constant 80 while each of the protein atoms is given a partial charge with a specific protein dielectric constant. From each grid, the electrostatic potential of the system is calculated iteratively based on equation 1.5 until converged and the electrostatic energy can then by evaluated by the effective potential acting the charges of each of the titratable group.

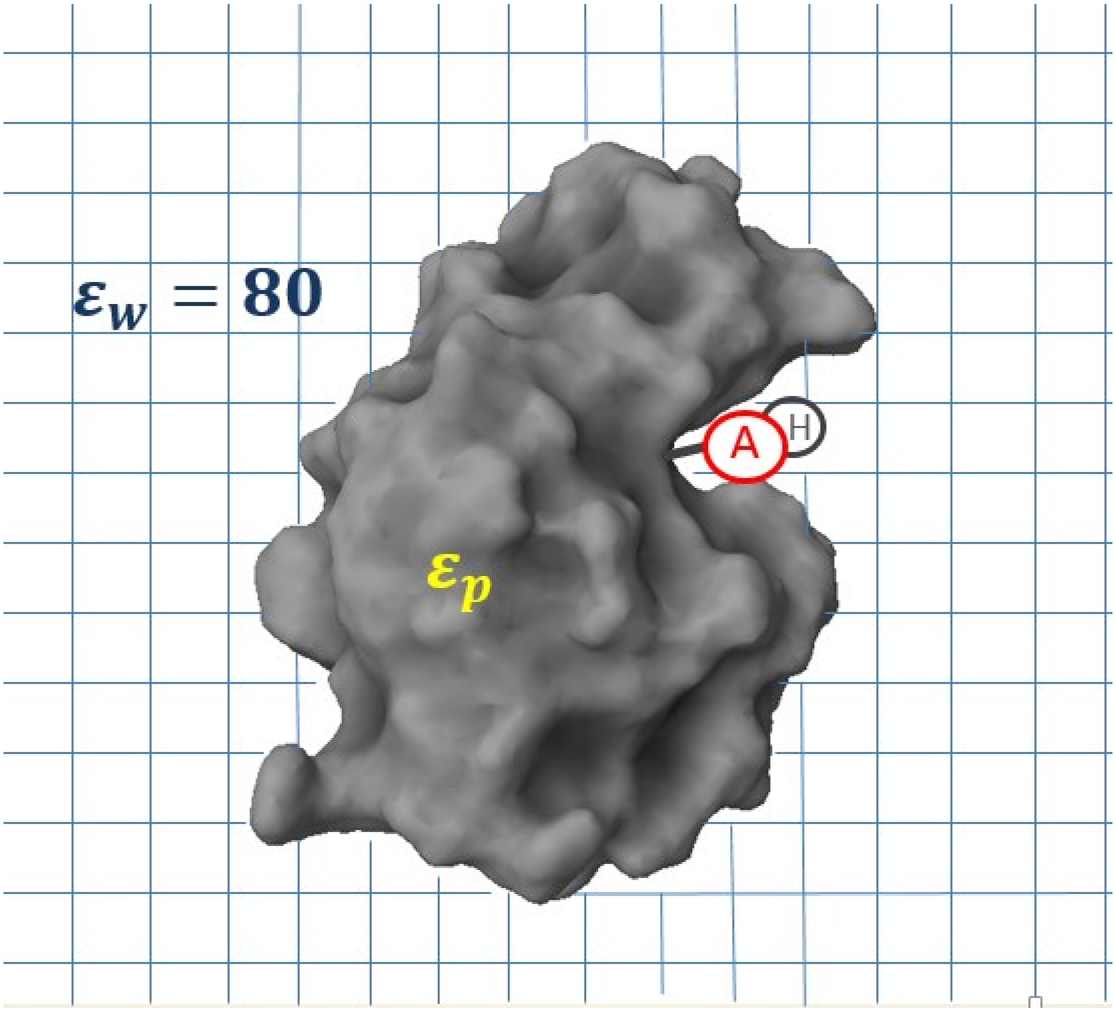


Figure 3. Poisson-Boltzmann electrostatic solvent models. The protein is implicitly treated as a dielectric medium. The dielectric constant of water, (ε_ω), is 80. The space is gridded up and each grid point represents a polarizable water molecule

1.4 Dielectric constant in protein

The major challenge in PB method is to determine the appropriate dielectric constant that best describes the heterogeneous protein environment. The meaning of the protein dielectric constant, ε_p, has been discussed repeatedly. [22-24] While early studies have assumed the protein dielectric constant as the experimentally determined protein dielectric constant, it is only recently realized that it is a simply scaling factor that accounts for missing electrostatic effects such as solvent reorientation and reorganization, protein flexibility, polarization effect, and other medium's responsiveness to charges within the electrostatic models. Thus, if an electrostatic model captures all the physically details of the described system in atomistic detail, the dielectric constant required for calculating all Coulombic interactions should be equivalent to 1. If an electrostatic model is described largely in a macroscopic way, such as neglecting the effect of protein relaxation and solvent reorganizations, the effective protein dielectric required to reproduce to experimental electrostatic behavior can be set as high as 10~20 to capture the missing dielectric screening effect due to the electrostatically induced solvent and protein reorganization. Therefore, the dielectric constant of protein depends on how the model describes the physical properties rather than directly being related to the experimental observations. In the recent meeting among the *pK_a-cooperative* members, a focus group working on current advances in pK_a calculation, it has been concluded that the best results generally could be produced with ε_p= 8~20 within the Poisson-Boltzmann model.[25-27] Unfortunately, while many implementations based on PB method reproduce the experimental pK_a quite well, they often fail to predict the pK_a of biological interesting and relevant ionizable groups that exhibit large pK_a shifts within buried sites. This has been pointed out earlier by Warshel and coworkers that the use of high dielectric constant leads the prediction to a null model where $\Delta pK_a = 0$ and even this null model would seem to predict pK$_a$ quite well because most of the protein pK$_a$ shifts are small.[23, 28] As a result, research focus on what protein dielectric constant should be used for accurate pK$_a$ prediction using the PB model remains an active area of research.

1.5 Evaluating protein pK$_a$ with PDLD/S Model: Intrinsic pKa and apparent pK$_a$

The semi-Microscopic Protein Dipole Langevin Dipole (PDLD/S) evaluates the electrostatic solvation free energy based on the LRA method.[23, 29] (See Appendix A for more details of LRA method) The approach adopted for protein pKa calculation involved a two steps approach that uses a single low dielectric constant for calculating the intrinsic protein pKa‘s when all other ionizable group are neutralized and a single large dielectric constant for evaluating the pKa‘s shifts as a result of charge-charge coupling between ionizable groups. The detail of the PDLD type model is described elsewhere. The two steps approach for evaluating the protein pKa is described as follows. To evaluate the intrinsic pKa, the self-energy of ionizing this group when all other ionizable groups are neutralized are decoupled from the charge-charge interaction within the protein, $\Delta G_{qQ}^{\mathrm{p}}$, and $\Delta\Delta G_{\mathrm{solv}}^{\mathrm{w}\rightarrow\mathrm{p}}$ can be expressed as

$$\begin{aligned}\left(\Delta\Delta G_{\mathrm{solv}}^{\mathrm{w}\rightarrow\mathrm{p}}\right)_i &= \Delta G_{q\mu}^{\mathrm{p}} + \Delta G_{q\alpha}^{\mathrm{p}} + \Delta G_{qw}^{\mathrm{p}} + \Delta G_{qQ}^{\mathrm{p}} - \Delta G_{\mathrm{self}}^{\mathrm{w}} \\ &= \left(\Delta G_{\mathrm{self}}^{\mathrm{p}} - \Delta G_{\mathrm{self}}^{\mathrm{w}}\right)_i + \Delta G_{qQ}^{\mathrm{p}} \qquad (1.6)\end{aligned}$$

.

Within this formalism, the charge-charge interactions are decoupled and are evaluated independently. The term, ‘self-energy’, is defined as the free energy associated with changing the charge of an ionizable group from zero to their average charge in its specific environment. It does not require the evaluation of the gas phase free energy as it is cancelled within the $\Delta\Delta G_{solv}^{w\rightarrow p}$. The self-energy term consists of the opposing energetic influences involving $\Delta G_{q\mu}^{\mathrm{p}}, \Delta G_{q\alpha}^{\mathrm{p}},$ and $\Delta G_{qw}^{\mathrm{p}}$ which are the free energy of the electrostatic interactions between the charge of an ionizable group and the surrounding permanent dipoles, polarizable dipoles, and water, respectively. Finally, equation 1.6 can be expressed in terms of $\mathrm{p}K_a$ as

$$\mathrm{p}K_{\mathrm{a},i}^{\mathrm{app}} = \mathrm{p}K_{\mathrm{a},i}^{\mathrm{int}} + \Delta\mathrm{p}K_{\mathrm{a},i}^{\mathrm{charges}} \qquad (1.7)$$

where $\mathrm{p}K_{\mathrm{a},i}^{\mathrm{app}}$ is the “expected” or the apparent $\mathrm{p}K_a$ of residue i in protein, $\mathrm{p}K_{\mathrm{a},i}^{\mathrm{int}}$ is the $\mathrm{p}K_a$ of i-th residue when all surrounding ionizable groups are neutralized and $\Delta\mathrm{p}K_{\mathrm{a},i}^{\mathrm{charges}}$ is the $\mathrm{p}K_a$ shift due to the charge-charge coupling between residue I and all surrounding ionizable residues.

The presence of ionized groups polarizes the local environment that can lead to large dielectric screening between charges. By evaluating the intrinsic $\mathrm{p}K_a$ when all other ionizable groups are neutralized, the approach focuses on evaluating the desolvation free energy associated with moving the ionizable group of interest from water to protein and circumvents the need to use of a large dielectric constant to properly describe the induce screening between the charged ionized groups. Implementation of this strategy into the existing PB model will be the main subject of my thesis. The work will focus on identifying the optimal protein dielectric constant for accurate protein pKa prediction.

1.6 Evaluating titration curve for monoprotic acid

Evaluation of charge-charge interactions have been introduced elsewhere. [23, 28, 29] To begin, one must first examine the ionization of the single amino acid side chain which is described by the proton dissociation reaction of a monoprotic acid.

$$\mathrm{AH} \rightleftharpoons \mathrm{A^-} + \mathrm{H^+}$$

Its Gibbs free energy of reaction is described by

$$\Delta G = -RT\ln K_a \tag{1.8}$$

where *R* is the gas constant, *T* is the temperature, and *K*a is the equilibrium acid dissociation constant defined as

$$K_\mathrm{a} = \frac{[\mathrm{A^-}][\mathrm{H^+}]}{[\mathrm{AH}]} \tag{1.9}$$

Such expression can be re-written in p*K*a units, $-\log(K_\mathrm{a})$, as the well-known Henderson-Hasselbalch equation

$$\mathrm{pH} = \mathrm{p}K_\mathrm{a} + \log\frac{[\mathrm{A^-}]}{[\mathrm{AH}]} \tag{1.10}$$

Denoting $\frac{[\mathrm{A}^{-}]}{[\mathrm{AH}]_0}$ as f_{acid} which is the fractional concentration of deprotonated acdis from its initial protonated state, Eq. 1.10 can be rewritten as

$$f_{\mathrm{acid}} = \frac{[\mathrm{A}^{-}]}{[\mathrm{AH}]_0} = \frac{1}{1 + 10^{(\mathrm{p}K\mathrm{a} - \mathrm{pH})}} \tag{1.11}$$

Now by multiplying the integer charge, q^0, of the acid(-1) or base(+1), the average charge of a given acid can be expressed

$$\langle q \rangle = \frac{q^0}{1 + 10^{\gamma(\mathrm{pH} - \mathrm{p}K_\mathrm{a})}} \tag{1.12}$$

where γ is +1 for base and -1 for acid. Note that at the point where pH is equal to $\mathrm{pK_a}$, the average charge <q> becomes 0.5. Therefore, by calculating Eq. 1.12 at each pH point whose interval is small enough to interpolate, we can find the apparent $\mathrm{pK_a}$ on its titration curve.

1.7 Evaluating interactions between titratable groups

The reaction free energy of deprotonation, ΔG^0 , for a monoprotic acid at a specific pH is given by

$$\Delta G^0 = -2.3RT\gamma[pKa - pH] \tag{1.13}$$

If pH around an acid is higher than its $\mathrm{pK_a}$, ΔG^0 is negative and the deprotonation reaction is spontaneous, and vice versa. Now, we need to consider the charge-charge interactions between *i*-th titratable residue with all other titratable groups. The charge-charge interaction can be evaluated within the macroscopic formalism using the Coulombic expression

$$\sum_{j\neq i}^{N} \Delta G_{ij}^{\mathrm{p}} = \sum_{j\neq i}^{N} \frac{\langle q_i\rangle\langle q_j\rangle}{r_{ij}\varepsilon_{ij}} = \sum_{j\neq i}^{N} \langle q_i\rangle\langle q_j\rangle W_{ij} \tag{1.14}$$

where r_{ij} is the distance and ε_{ij} is the effective dielectric constant (normally denoted as ε_{eff} is a single uniform dielectric is used) between *i*-th and *j*-th ionized residues. Because ε_{ij} involved interaction between charges and is described with in a macroscopic way, the ε_{eff} of 40 and higher can be used.

<q> is the effective average charges at the given pH evaluated based on eq. 1.12 . The total free energy of the *i*-th residue is given by combining equation 1.13 and 1.14 as

$$\Delta \mathrm{G_i} = \Delta \mathrm{G}^0 + \Delta \mathrm{G_{ij}^p}$$

$$= -2.3RT\gamma[pKa - pH] + \sum_{j\neq i}^{N} \langle q_i\rangle\langle q_j\rangle W_{ij} \tag{1.15}$$

while the average charge of *i*-th residue, $<q_i>$, can also be defined as [23]

$$\langle q_i\rangle = \frac{q_i^0 \exp^{-\beta\Delta G_i}}{\left(1+\exp^{-\beta\Delta G_i}\right)} \tag{1.16}$$

where β is the inverse of the thermodynamic temperature and q_i^0 is the initial charge of the titratable group, thus, +1 and -1 for base and acid respectively. Note that equation 1.15 and 1.16 are solved self-consistently through iteration until converged. By titrating the average charge, the apparent pK_a is found where <q> is equal to ±0.5 based on Eq. 1.16.

1.8 Project Objectives

Although Warshel's group has repeatedly pointed out the advantages of this decoupling charge-charge interactions strategy, this still has been overlooked and misunderstood in PB models. In the meanwhile, the reliability of pK_a calculations has not been improved as much as the development of new and complicated methods with the increase of computational power over the past decades. [25] Interestingly, to the best of our knowledge, there is no Poisson-Boltzmann based method that has been described to–date that has adopted this two steps method by first evaluating the intrinsic pK_a and then the pK_a shift from charge-charge interactions using two different dielectric constants

In this thesis, the previously described PDLD/S approach is incorporated into Poisson-Boltzmann model to verify the hypothesis that both general and large pK_a shifts can be more reliably predicted by decoupling charge-charge interactions and applying more consistent small dielectric constant for intrinsic pK_a calculation. Our results are compared to the classic PB method and other benchmarks to identify the optimal dielectric constants required by the model. Finally, we will demonstrate that incorporating protein relaxation within this new approach can further improve its predictive pKa accuracy.

Methods

2.1 Implementation of intrinsic pK$_a$ calculation

A general process flow chart is given in Figure 4. The details of the implementation is described below.

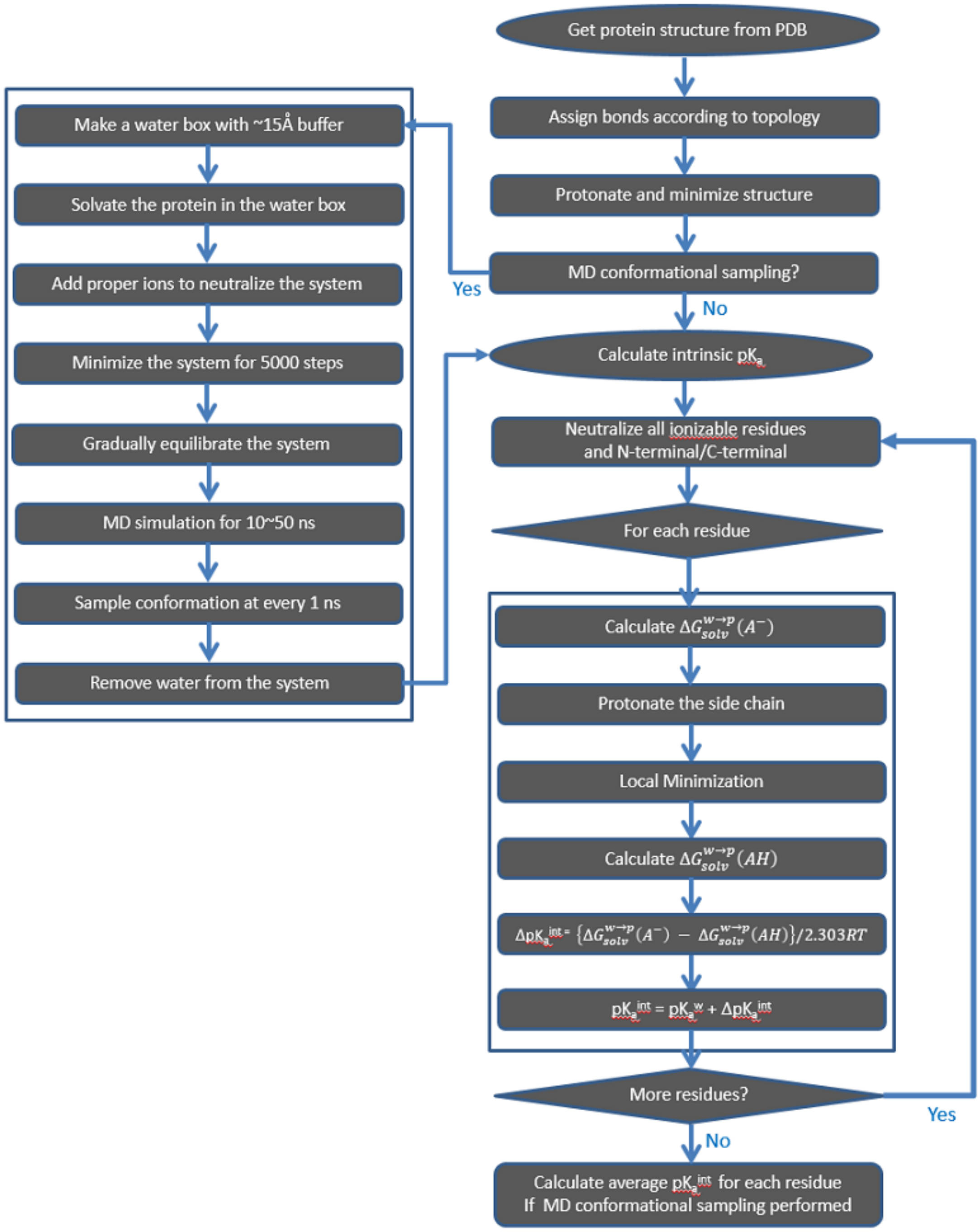

Get protein structure from PDB
Assign bonds according to topology
Protonate and minimize structure
MD conformational sampling?
Yes
No
Make a water box with ~15Å buffer
Solvate the protein in the water box
Add proper ions to neutralize the system
Minimize the system for 5000 steps
Gradually equilibrate the system
MD simulation for 10~50 ns
Sample conformation at every 1 ns
Remove water from the system
Calculate intrinsic pK$_a$
Neutralize all ionizable residues and N-terminal/C-terminal
For each residue
Calculate $\Delta G_{solv}^{w\rightarrow p}(A^-)$
Protonate the side chain
Local Minimization
Calculate $\Delta G_{solv}^{w\rightarrow p}(AH)$
$\Delta pK_a^{int} = \{\Delta G_{solv}^{w\rightarrow p}(A^-) - \Delta G_{solv}^{w\rightarrow p}(AH)\}/2.303RT$
$pK_a^{int} = pK_a^{w} + \Delta pK_a^{int}$
More residues?
Yes
No
Calculate average pK$_a^{int}$ for each residue If MD conformational sampling performed

Figure 4 The flow chart of intrinsic pK$_a$ calculation

2.1.1 Protein preparation

X-ray protein crystal structures are obtained from Protein Data Bank. CHARMM (Chemistry at Harvard Molecular Mechanics), a program for macromolecular simulations, [14] is used for structure preparation. CHARMM27 protein force field is used for atomic description of the protein structure. [30] Since original X-ray structures do not include hydrogen and disulfide bonds, they are assigned systematically based on the CHARMM27 force field.

2.1.2 Electrostatic free energy calculations

pK$_a$ calculations based on original X-ray crystallographic structure are taken directly after protein preparation in 2.1.1. PKa calculation that account for protein relaxation is evaluated using the trajectory obtained from MD simulation. The snapshots of each protein structure coordinates taken at every 1 nanosecond as independent structural conformation sample were used with water and counter ions surrounding the proteins removed. The intrinsic pK$_a$ calculation was implemented with CHARMM script language. For each protein structure conformation (X-ray or simulated MD snapshot), all ionizable residues (ASP, GLU, TYR, SER, LYS, ARG, and HIS) as well as the N-terminal and C-terminal regions are neutralized by reassigning partial atomic charges that can reproduce the experimental solvation energy of that chemical entity. The pK$_a$ calculation is carried out iteratively for each ionizable residues based on the PB model implemented within CHARMM.

For each PB calculation, the dielectric constants for the protein interior and water are set to 4 and 80, respectively. 0.15M of salt concentration is used. The grids are generated around the protein with 1.5 Å spacing. For better accuracy, smaller spacing with 1 Å is applied around the indicated ionizable group. The electrostatic free energy of the indicated group in deprotonated states from water to protein is calculated by Poisson-Boltzmann equation module in CHARMM. Now the side chain of the indicated group is protonated and all hydrogen positions within 4 Å of the group are energy minimized to ensure that the added hydrogen does not sterically clash with others atoms. The electrostatic free energy of the group in protonated states in water and in protein site calculated in the same way with the same parameters. The intrinsic pK_a shifts are calculated based on Eq. 1.9. For pKa calculation with MD simulation, these steps are “embarrassingly” parallelized with scripting by distributing over large number of serial processes and repeated for all sampled conformations. The intrinsic pK_a values for each ionizable residue along the trajectory are averaged based on Linear Response Approximation (LRA) method.

2.1.3 Molecular Dynamics simulation

Each protein structure is solvated in a box with TIP3P explicit water model[31] with 15 Å buffer region from the surface of the protein structure. Na^+ or Cl^- ion is added at 2 Å from the box boundary to electroneutralize the total charge of the system. MD simulation is performed using NAMD version 2.6. [32] with periodic boundary condition using Particle Mesh Ewald (PME) [33]. Each system is energy minimized with conjugate gradient algorithm for 5000 steps with 50kcal/(mol•Å^2) restrain on each heavy atom. SHAKE method [34] was employed allowing only hydrogen atoms to move at fixed bond length. During initialization the restraint system is gradually heated from 25 K to 300 K increasing 25 K at every 10 picoseconds for 100 picoseconds at 2 femtosecond time step. For the next 100 ps, the heavy atom restraints are gradually decreased and removed under NVT condition. The final unrestrained equilibration is carried out for 100 ps followed by 10~50 nanoseconds of MD simulation at 1atm and 300K under NPT condition. Snapshots of the protein-water system coordinates are saved at every 1 picosecond. If the simulation is successfully finished, the configurations along the trajectory is superimposed to the initial structure and the divergence and the stability of the protein structure is evaluated with Cα atoms roots mean square deviations (RMSD) plot generated from the RMSD trajectory tool in VMD. [35] If RMSD shows large structural fluctuations a small constrain is employed during the simulation. All MD simulation were carried out using Itasaca high performance computer at the University of Minnesota Supercomputing Institute.

2.2 Calculation of pK_a shifts by Charge-Charge interaction

An overall procedure is given in Figure 5. The detail of the implementation is described as follows.

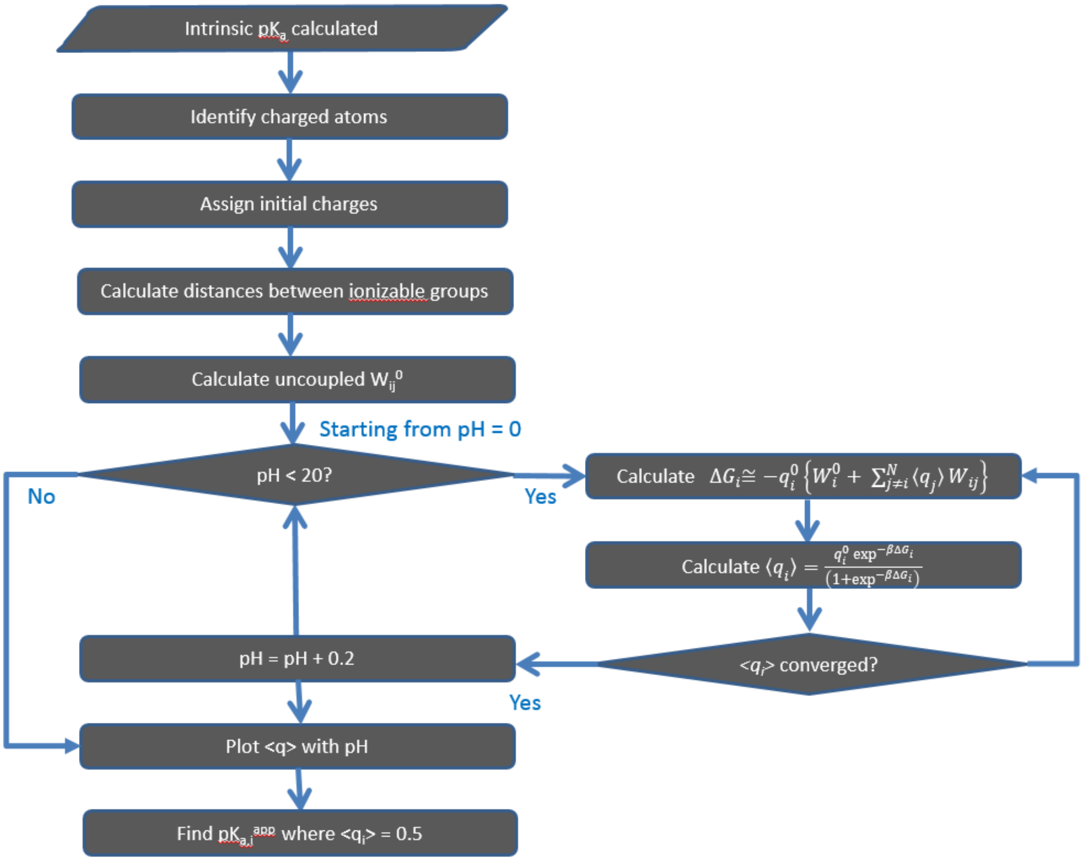

Intrinsic pK_a calculated
Identify charged atoms
Assign initial charges
Calculate distances between ionizable groups
Calculate uncoupled W_{ij}^0
Starting from pH = 0
pH < 20?
No
Yes
Calculate $\Delta G_i \cong -q_i^0 \left\{W_i^0 + \sum_{j\neq i}^{N} \langle q_j \rangle W_{ij}\right\}$
Calculate $\langle q_i \rangle = \frac{q_i^0 \exp^{-\beta \Delta G_i}}{(1+\exp^{-\beta \Delta G_i})}$
No
<q_i> converged?
Yes
pH = pH + 0.2
Plot <q> with pH
Find $pK_{a,i}^{app}$ where <q_i> = 0.5

Figure 5 The flow chart of apparent pKa calculation

2.2.1 Preparation

Once the intrinsic pKa is calculated, the values are used as the starting points to evaluate the apparent pKa based on equation 1.15 and 1.16. The module was developed in Perl. For the macroscopic treatment of the charge-charge interaction, the protein is treated as a macroscopic medium of large dielectric constant with only the ionizable side chains are considered. Each ionizable side chain is assigned either a single or double ionized centers based on the chemical nature of the ionizable group as shown in Table 2. For Ser, Tyr and Lys, a +/- 1 charge is assigned to the single electronegative atom as the ionizable center. For Arg, Lys, His, Asp and Glu, double ionizable centers are assigned with an initial value of +/- 0.5 charge to reflect on the multiple tautomeric protonation states of the side chain.

Residue	Atom Type	Initial charge
Arginine	NH1&NH2	0.5 for each
Lysine	NZ	1
Histidine	ND1&NE2	-0.5 for each
Aspartate acid	OD1&OD2	-0.5 for each
Glutamic acid	OE1&OE2	-0.5 for each
Tyrosine	OH	-1
Serine	OG	-1

Table 2 Atom types that used to calculate the charge-charge interactions energy and the initial charges assigned

As shown in Eq. 1.11, the distance between two charges is one of the factors used to evaluate charge-charge interactions. The computational complexity of calculating all the distances in the system is O(n!) where n is the number of the ionizable site. Therefore, the computation cost would be dramatically increased as the number of ionizable sites increases. Moreover, the charge-charge coupling calculation is an iterative procedure over the incremental range of pH. Therefore, the total computational cost becomes O(n! x n!). To improve on the overall efficiency, the interatomic distance, r_{ij} is calculated only once for each protein structure conformation in the beginning and stored as a lookup two dimensional matrix table for the iterative Coulombic interaction energies calculation.

2.2.2 Calculation of Charge-Charge interaction energies

The charge-charge interaction energies are calculated based on Eq. 1.16 and 1.17 at 0.2 pH intervals. Uncoupled free energy, W_i^0 for each titratable group is calculated at incremental pH based on the calculated intrinsic pK_a. The Eq. 1.18 and 1.19 are solved iteratively until convergence is achieved. The Coulombic energy between residue i and j, $\frac{q_i \ q_j}{r_{ij}\varepsilon_{eff}}$, is stored in an n x n matrix where n is the total number of the sites. For those residues whose side chains have two protonation sites, the interaction energy is calculated for both atoms and summated as one. For example, when the charge-charge interaction energy between a lysine and a glutamic acid group is calculated, two interactions are considered between NZ and OE1, and between NZ and OE2. By assigning a half of the charge to each atom, we can reflect the resonance form more consistently.

2.2.3 Titration curve and apparent pKa

Once the average charges, $<q_i>$, for the titrated for each residue is evaluated, the pH point where $<q_i>$ becomes a half of its initial charge (+1 for base. -1 for acid) is identified as apparent pK_a. (Figure 6). Because it is titrated at 0.2 pH unit intervals, there is no guarantee that one of the titration point will exactly hit $<q_i>=0.5$. Because calculating with smaller intervals increases the computational cost, it is better to approximate the apparent pK_a point assuming that the titration curve around $<q_i>=0.5$ is almost linear. The titration curve is traced from the both sides until the closest upper bound and the lower bound from 0.5 are found. The apparent pK_a is calculated using the linear properties with the proportions as described in Figure 6.

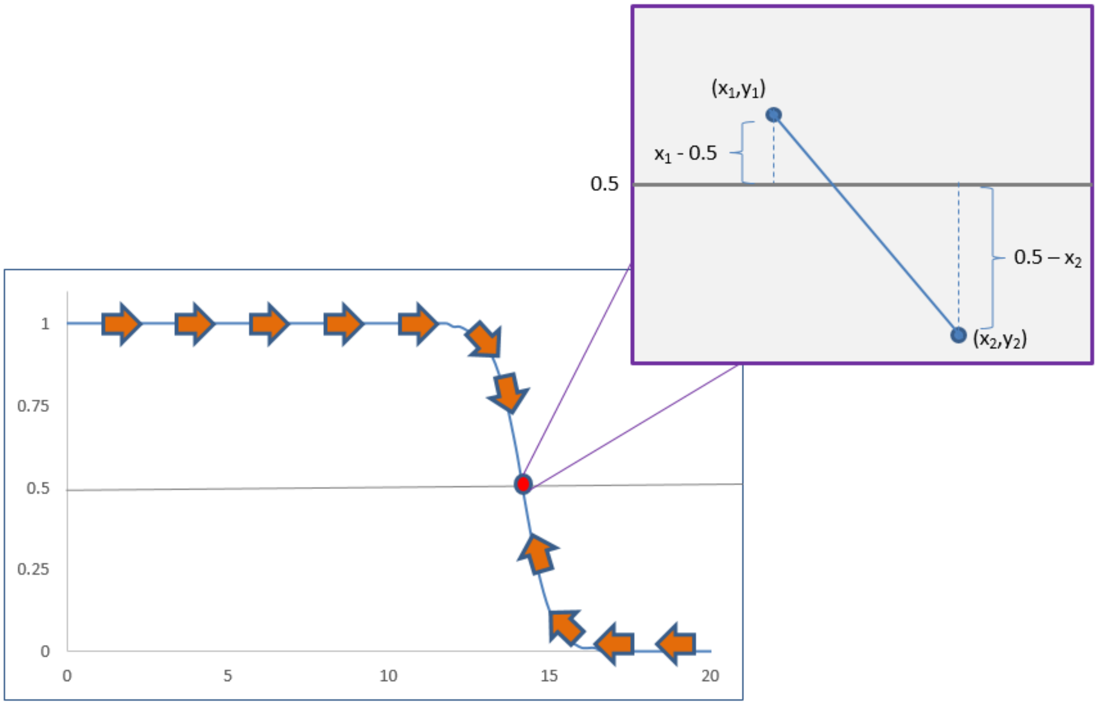


Figure 6 A description showing how to find apparent pK$_a$ which is at <q$_i$>=0.5. The curve is traced from both sides and the upper and lower bounds closest to 0.5 are selected. Assuming that the curve is close to linear, the midpoint calculated using $\frac{(b-0.5)}{(x2-x1)}$ ×(x2-x1) + x1

2.2.4 Simple case tests

Two simple cases were tested to see if this program captures the correct pK$_a$ shift trends. (Figure 7). Two ionizable groups were put together within 4 Å. In principle, the interaction between opposite charges are energetically favorable. As seen in Figure 7-(a), the base and acid stabilize each other shifting pK$_a$ down for the acid and shifting pK$_a$ up for the base to stay in charged states. In contrast, two bases nearby destabilize each other and tend not to have a charge-charge interaction by shifting down the pK$_a$.(Figure 7-(b)) Note that there is no shift in the titration curve for the lysine group. This can be explained by the fact that once the charge of the histidine is zero, there shouldn’t be any coupling.

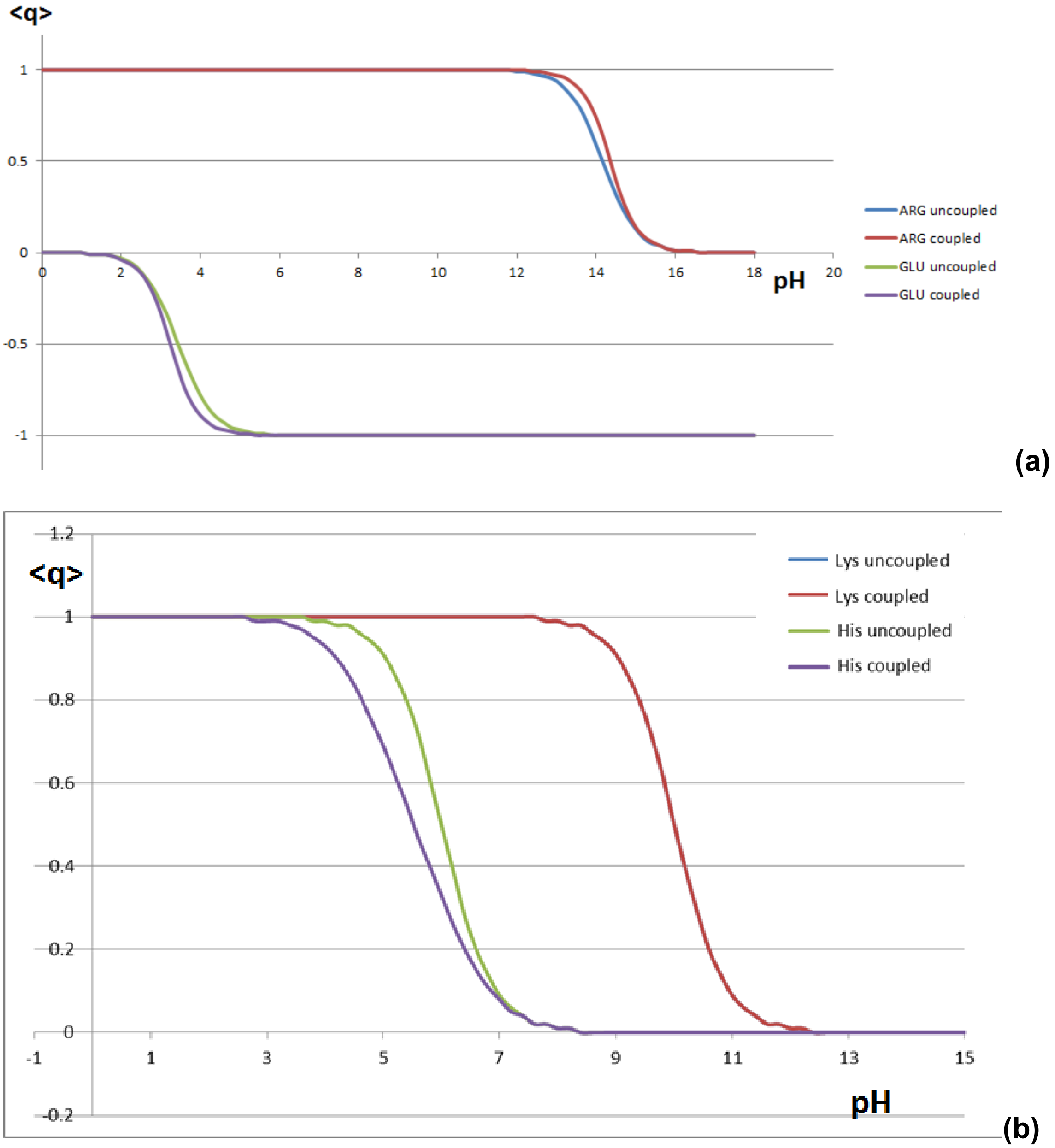


Figure 7 Simple test cases. Two ionizable groups were located 4 Å away from each other and the average charges were titrated.

2.3 Protein pKa Benchmark

To evaluate our model, 7 protein structures were used with their well-established experimental pK_a data. (Table 3)

Proteins	Number of residue	PDB	Experiment set ref
Lysozmye	26	1HEL	[36]
RNaseA	28	7RAA	[37]
Ovomucoid	11	1OMU	[38, 39]
Barnase	28	1A2P	[40, 41]
Thioredocxin-Oxidized	30	1TRS	[42]
Thioredoxin-Reduced	30	1TRW	[42]
BPTI	14	5PTI	[43]

Table 3 Protein structures tested and their experimental pK_a data references. MD simulations were performed for each protein for 20ns.

3. Results and Discussion

3.1 PB method with single dielectric constants

First, we tested the dependence of the dielectric constants in the classic PB method that uses charged states with single dielectric constants and single structures. The comparisons between calculated and experimental pK_a values are listed in Table 4. We tested a set of dielectric constants, 4, 6, 10, and 20. At this point, our focus was on the effect of the dielectric constants in classical PB model. Thus, single structures are used for predictions. The numbers listed in Table 4 are root mean square deviations from the experimental pK_a.

It is clear from the results that when a higher dielectric constant is used, the numbers show better correlations with the experimental values in overall. However, the good correlations observed at high dielectric constants such as ε_P= 20 or 40 do not necessarily mean that the model is consistent. As pointed out earlier, when compared to the null model, where $\Delta pKa = 0$, the PB model reproduces similar results at ε_P = 10 and 20. This result gives rise to the uncertainty of whether the best correlation at ε_P = 20 and 40 is due to the lack of sensitivity of the electrostatic model itself. The use of a high dielectric constant screens significantly the electrostatic interaction with its surrounding environment which can inadvertently leads to the null model outcome. As most of the pK_a shifts are experimentally observed to range below 1 pK_a unit, it become important to question the consistency of the given model even when the RMSD appears "seemingly" more accurate. The prediction for ovomucoid is the only exceptional case from this trend. Unlike other proteins whose most of the experimental pK_a shift from water pK_a are less than 1.0 pK_a unit, half of the residues in ovomucoid showed more than 1.2 unit of experimental pK_a shift from water pK_a. Therefore, merely decreasing overall pK_a shifts with high dielectric constant in ovomucoid results in an opposite trend which predicts pK_a further from the experimental pK_a.

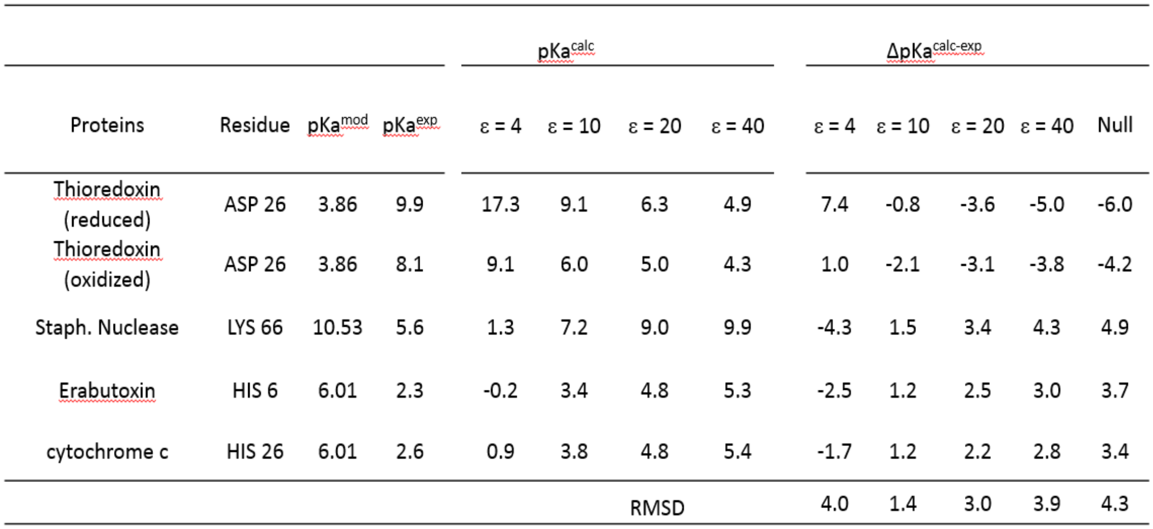
Proteins	Number of Residue	ε = 4	ε = 6	ε = 10	ε = 20	ε = 40	Null Model
Lysozmye	10	2.6	1.9	1.4	1.1	1.1	1.5
RNaseA	14	2.3	1.6	1.1	0.7	0.5	0.7
Ovomucoid	6	0.6	0.6	0.7	0.8	0.9	1.1
Barnase	10	3.8	2.6	1.7	1.0	0.8	1
Thioredocxin-Ox	17	1.5	1.1	1.0	1.0	1.1	1.2
Thioredoxin-red	17	2.4	1.4	0.9	1.1	1.4	1.6
BPTI	13	1.0	0.9	0.8	0.8	0.7	0.7
Total	87	2.3	1.6	1.3	1.0	1.0	1.2

Table 4 Classic PB method in function of single dielectric constant. Listed numbers root mean square deviations (RMSD from experimental pK_a values. The null model represents when there is no pK_a shift which is $pK_a^{mod} – pK_a^{exp}$.

One may argue that using high dielectric constants by decreasing the overall pK_a shift still predicts pK_a values close to experimental data well. Indeed, Antosiewicz *et al.* [27] and Teixeira *et al.* [26] concluded that using single dielectric constant at 20 generate reliable results in PB method. One way to evaluate the possible false positive is to test special cases whose experimental pK_a shifts are large. Warshel and his coworkers tested their PDLD/S model to predict such discriminative pK_a shifts in their previous work by decoupling charge-charge interactions. [29]. Here, we performed the similar experiment but with PB model. The classic PB method with single dielectric constants was used to calculate pK_a of the residues that are well known for their experimentally observed huge pK_a shifts. (Table 5)

As expected, the use of high single dielectric constant of 20 and 40 severely underestimated the pK_a shifts for these residues. The average pK_a shift from the model pK_a at ε_P= 20 is 1.4 which is much smaller than that of the null model (4.3 pK unit). In contrast, the use of ε_P= 4 overestimates the large pK_a shifts observed in experiment. This can be especially observed in the calculated results of the the pK_a of HIS6 of erabutoxin b which was estimated even below zero.

Optimal results were observed when ε_P= 10 was used to predict the pK_a. This predicts general cases quite well while also predicting large pK_a shifts with smaller errors. These results coincide with the general agreement from the meeting among the *pK_a cooperative* members, a focus group of researchers working on pK_a predictions. They observed that a majority of PB based methods usually generate the best results at ε_P= 8 to 10 although they still saw significant errors occasionally.[25] Starting from this classic method, we introduce our method that decouples charge-charge interaction in the next section.



				pK_a^{calc}				$\Delta pK_a^{calc-exp}$				
Proteins	Residue	pK_a^{mod}	pK_a^{exp}	ε = 4	ε = 10	ε = 20	ε = 40	ε = 4	ε = 10	ε = 20	ε = 40	Null
Thioredoxin (reduced)	ASP 26	3.86	9.9	17.3	9.1	6.3	4.9	7.4	-0.8	-3.6	-5.0	-6.0
Thioredoxin (oxidized)	ASP 26	3.86	8.1	9.1	6.0	5.0	4.3	1.0	-2.1	-3.1	-3.8	-4.2
Staph. Nuclease	LYS 66	10.53	5.6	1.3	7.2	9.0	9.9	-4.3	1.5	3.4	4.3	4.9
Erabutoxin	HIS 6	6.01	2.3	-0.2	3.4	4.8	5.3	-2.5	1.2	2.5	3.0	3.7
cytochrome c	HIS 26	6.01	2.6	0.9	3.8	4.8	5.4	-1.7	1.2	2.2	2.8	3.4
						RMSD		4.0	1.4	3.0	3.9	4.3

Table 5 Reported large experimental pK_a shifts and the deviations of calculated pK_a from experimental pK_a using classic PB model that uses charged states. The experimental pK_a were determined in [44], [45], and [46] for staphylococcal nuclease, erabutoxin b, and horse heart cytochrome c, respectively. The PDB structures used here are 2SNM for Staphylococcal nuclease, 3EBX for erabutoxin b, and 1HRC for horse heart cytochrome c.

3.2 Results of decoupling charge-charge interaction with PB method

The previous section addressed the challenge of the classic PB method. The model, while valid in a number of situations, is inadequate in addressing large pK_a shifts which are biologically important and relevant. As such, the challenge of this study is to see if we can develop a method in which we can avoid the dependence on the dielectric constant.

Sham *et al.* pointed out that it is possible to examine self-energy and charge-charge interaction independently by decoupling these two terms.[29] By doing so, one can use a more consistent dielectric constant for the intrinsic pK_a

while the use of high dielectric constant for charge-charge interaction (W_{ij}) is allowed. However, their work was implemented on PDLD/S which is a semi-microscopic model. Here, we tested this approach using PB model. We separated the charge-charge interaction terms with ε_{eff} = 40 within 15Å from the site and 80Å for other residues outside this range. The intrinsic pK_a were calculated when all ionizable residues are neutralized based on Eq. 1.19. The same set of dielectric constants were used for the same proteins. The calculated apparent pK_a are listed in Table 6. Compared to the classic PB method, slight improvements are shown in most of the cases. Unlike the results from the classic PB method that show the best correlation with experimental pK_a at ε_P= 10, a better correlation was always observed at higher dielectric constants, thus, the best results were obtained at ε_P= 20.

Although the overall accuracy did not change much, there are significant improvements in the predictions for lysozyme and RNaseA. In lysozyme, half of the calculated pK_a showed large errors ranging 1.5~5.9 units at ε_P= 4 in the classic PB method. As a result, with our method, large improvements were observed in a majority of the ionizable residues.

However, there were several residues for which both the classic PB method and our method could not account for. For example, GLU7 still showed a large error of 2.4 unit at low dielectric constants with our method. Additionally, in barnase, the predictions at ε_P= 4 with both classic PB method and our method were still very different from the experimental pK_a although most of the numbers were improved after treating intrinsic pK_a and charge-charge couplings separately. Especially, extremely large errors were observed for GLU73 and GLU75. When the intrinsic pK_a and charge-charge interactions are inspected separately, it can be observed that the intrinsic pK_a shifts mostly accounted for the large pK_a shifts obtained. Therefore, these large perturbations must not be from charge-charge interactions, but rather caused by unconsidered protein relaxation effects or inappropriate dielectric boundary conditions since they are located on the surface.

For a better overview, all predicted pK_a values were plotted against the experimental data in Figure 8. Overall, the plots show that most of these ionizable groups benefits from higher dielectric constants. While significantly large errors such as ASP26 in thioredoxin are observed, the model is shown to be accurate in a majority of residues observed. However, as seen in calculations with the classic PB method, this does not necessarily mean that the prediction is consistent because it may fall into a null model. To see if our model still shows the null model trend as the classic PB method, plots of the null model versus our calculated pK_a shifts are shown in Figure 9. Note that most of the plots for the null model are ranged between -1 and 1. These plots clearly show that most of the predicted pK_a shifts became smaller when the dielectric constant was increased and as a result, the plots become flatter which reflects a lower consistency approaching the null model. Again, this makes the results look highly correlated with the experimental pK_a by forcing most of the predictions to be in the similar range as pK_a shifts in the null model. The problem is that the model that uses a high dielectric constant also scales down all other large shifts within this range. To use high dielectric constant for good overall prediction, it is inevitable to give up the accuracy for biologically important or relevant residues that have large pK_a shifts.

For a better insight on the effect of the charge-charge couplings, the intrinsic pK_a and W_{ij} at ε_P= 4 and 10 are listed in Table 7. Our predictions for most of the proteins were improved by adding W_{ij} to the intrinsic pK_a. However, the apparent pK_a's were not perturbed much by W_{ij} which shows around 0.8 shifts from the intrinsic pK_a in average. Our results correspond to a mutagenesis study in which small effects on pK_a by charge charge interactions were observed by testing how much pK_a is changed by mutating target ionizable residues to nonpolar residues. Their results found that there were only about 1 pK_a unit changes.[47] This is because there should be large dielectric screenings between charge-charge interactions. Indeed, our results show that when a ε_{eff}= 40~80 is used, the observed W_{ij} remains within a reasonably small range.

Our method so far has shown that a similar or even better predictions can be achieved with the strategy of decoupling charge-charge interactions. Now, as our main focus in this study, the calculations for the large experimental pK_a with our method were conducted and the results are listed in Table 8. The intrinsic pK_a shifts at ε_P= 4 and the W_{ij} with ε_{eff}= 40,80 are listed in Table 9 for better insight of the effect of decoupling charge-charge interactions. In contrast to the classic PB method where ε_P= 10 generated the optimal results, the accuracy of our predictions is always seen better at lower dielectric constants. This trend is exactly opposite to the results from general cases seen in Table 6 with our methods. This reflects the dilemma more clearly that even though we get better results in general with higher dielectric constants approaching the null model, we really need to use low constants to accurately predict such large pK_a shifts. However, using single dielectric constants means that dielectric screening from both charge-charge interaction and other induced dipole or non-polar interactions are adjusted at the same extent. Therefore, this leads to overestimation of charge-charge interactions at low dielectric constants and underestimation at high dielectric constants as observed in Table 6. Even though the best prediction from the classic PB method in Table 5 could be obtained at ε_P= 10 for the large pK_a shifts, errors larger than 1.2 unit were observed in all of the calculations. In our model, we could solve this dilemma by decoupling the charge-charge couplings and predict these large pK_a shifts using ε_P= 4 as accurate as the classic PB model that used ε_P= 10. This strongly supports our motivation in this study.

Proteins	PDB	ε = 4	ε = 6	ε = 10	ε = 20
Lysozmye	1HEL	1.3	1.0	1.0	0.9
RNaseA	7RAA	1.5	1.2	0.9	0.8
Ovomucoid	1OMU	0.8	0.7	0.6	0.6
Barnase	1A2P	3.6	2.5	1.6	1.1
Thioredoxin-ox	1TRS	1.2	1.1	1.2	1.2
Thioredoxin-red	1TRW	2.4	2.0	1.7	1.8
BPTI	5PTI	0.9	0.5	0.3	0.3
Total		1.9	1.5	1.2	1.1

Table 6. RMSD of calculated apparent pK_a with our method a dielectric constants of 4,6,10, and 20 for intrinsic pK_a calculation when all ionizable residues are neutralized.

Proteins	Intrinsic	Apparent
Lysozmye	2.0	1.7
RNaseA	1.7	2.1
Ovomucoid	1.3	0.8
Barnase	3.5	3.6
Thioredocxin-Ox	1.3	1.5
Thioredoxin-red	2.1	2.5
BPTI	1.5	0.9
Total	2.0	1.9

Table 7 RMSD of calculated intrinsic and apparent pK_a before and after W_{ij}.

				pKa^{calc}			$\Delta pKa^{calc-exp}$		
Proteins	Residue	pKa^{mod}	pKa^{exp}	ε = 4	ε = 10	ε = 20	ε = 4	ε = 10	ε = 20
Thioredoxin (red)	ASP 26	3.86	9.9	10.2	6.7	5.2	0.3	-3.2	-4.7
Thioredoxin (ox)	ASP 26	3.86	8.1	9.3	5.5	4.5	1.2	-2.6	-3.6
Staph. Nuclease	LYS 66	10.53	5.6	4.1	7.5	9.2	-1.5	1.9	3.5
Erabutoxin	HIS 6	6.01	2.3	2.1	3.0	4.5	-0.2	0.8	2.3
cytochrome c	HIS 26	6.01	2.6	2.3	0.8	1.4	-0.3	-1.9	-1.2
						RMSD	0.9	2.2	3.2

Table 8 Large experimental pK_a shifts and calculated pK_a for those residues when charge-charge interactions are decoupled by neutralizing all other ionizable residues. Refer to Table 6 for the detail reference for experimental values.

Proteins	Residue	ΔpKa^{exp}	ΔpKa_{int}^{calc}	W_{ij}	ΔpKa_{app}^{calc}
Thioredoxin (r)	ASP 26	6.0	5.5	0.8	6.3
Thioredoxin (o)	ASP 26	4.2	4.7	0.7	5.4
Staph. Nuclease	LYS 66	-4.9	-6.2	-0.2	-6.4
Erabutoxin	HIS 6	-3.8	-1.3	-2.7	-4.0
cytochrome c	HIS 26	-3.4	-3.3	-0.4	-3.7

Table 9 Large experimental pK_a shifts vs intrinsic pK_a shifts, charge-charge interaction term(W_{ij}), and apparent pK_a shifts

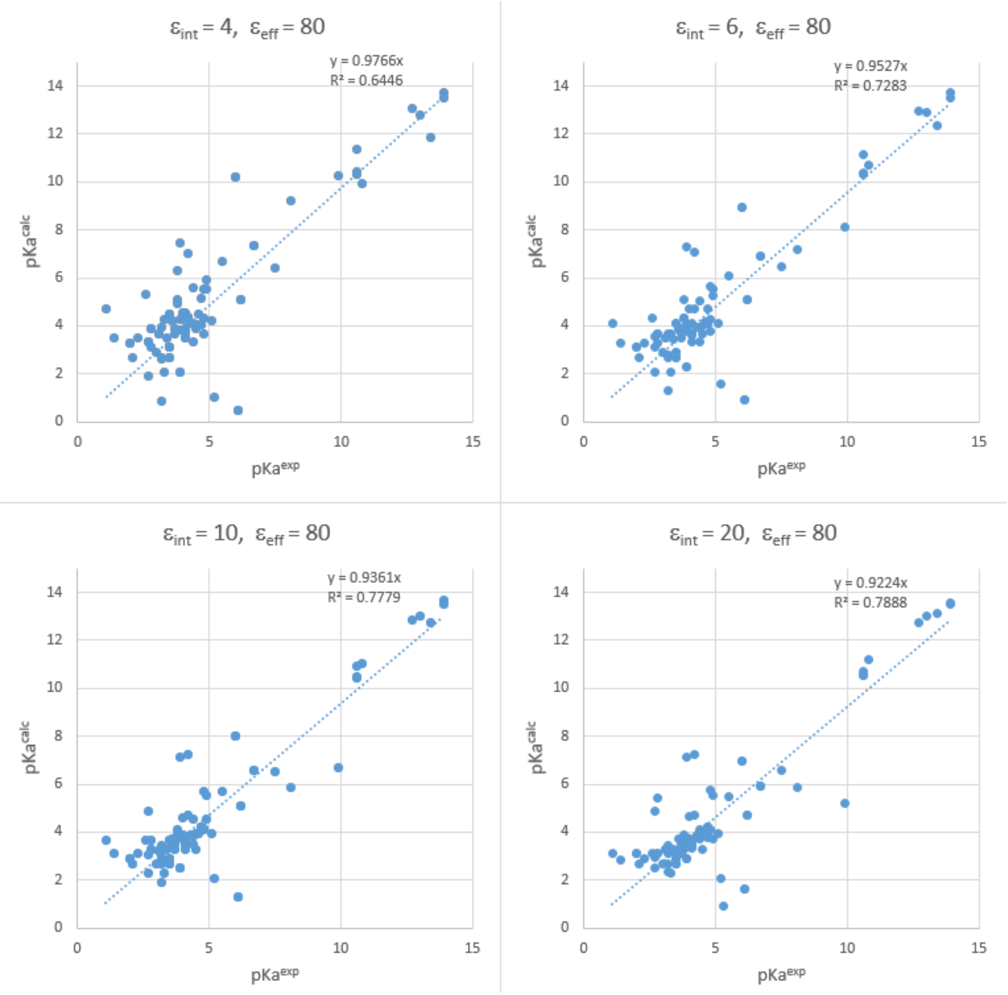


Figure 8 Experimental vs calculated pK_a for all 87 sites from the 7 proteins using four different dielectric constant, 4,6,10, and 20.

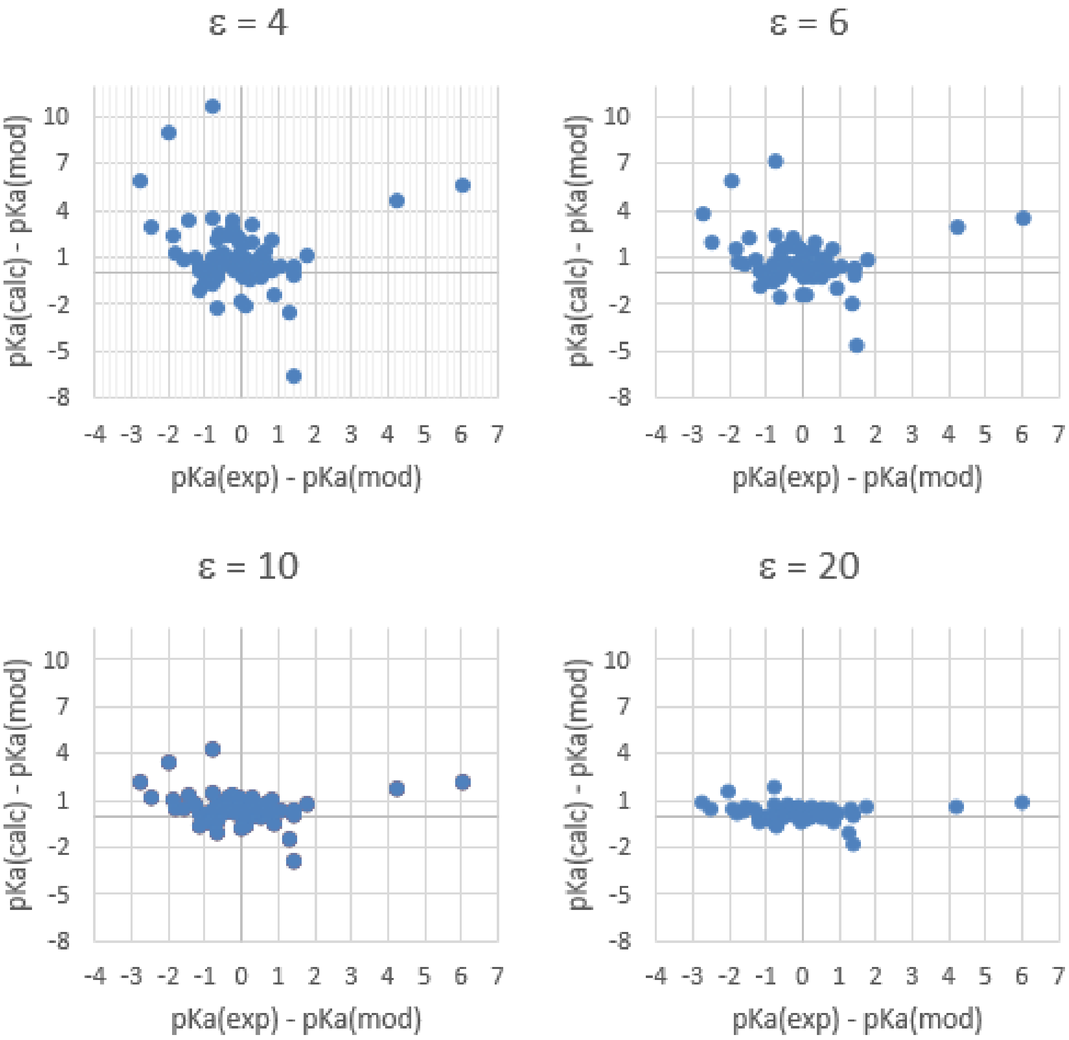


Figure 9 Experimental vs calculated pK$_a$ shifts from model pK$_a$ for all 87 residues from 7 proteins using different

3.3 Statistical sampling of conformations

Even though the results so far show that we can effectively predict the discriminative large pK$_a$ shifts at a low dielectric constant by decoupling charge-

charge interactions, the overall accuracy is not satisfying with ε_P= 4 for intrinsic pK_a calculation. One important factor that has not been addressed so far is the effect of protein relaxation. We performed a MD simulation for each protein and took conformations every 1 ns to get statistically independent samples. All individual calculations are listed in Table 10 after averaging the pK_a calculations at ε_P= 4 and the overall summaries are listed in Table 11.

Residue	pK_a^{int}	W_{ij}	pK_a^{app}	pK_a^{exp}	$\Delta pK_a^{calc-axp}$	Residue	pK_a^{int}	W_{ij}	pK_a^{app}	pK_a^{exp}	$\Delta pK_a^{calc-axp}$
Lysozyme						**BPTI**					
ASP18	3.4	-1.4	2.0	2.7	-0.7	Asp3	4.2	-0.7	3.5	3.4	0.1
ASP48	1.7	-1.3	0.5	2.5	-2.0	Glu7	7.0	-1.1	5.9	3.8	2.1
ASP52	4.2	-0.7	3.5	3.7	-0.2	Lys15	10.6	-0.2	10.4	10.6	-0.2
ASP66	1.9	-1.2	0.7	2.0	-1.3	Arg17	11.6	0.1	11.7	12.7	-1.0
ASP87	3.1	-1.1	2.1	2.1	0.0	Arg20	12.2	0.3	12.5	13.9	-1.4
ASP101	4.0	-1.0	3.0	4.1	-1.1	Lys26	10.2	-0.1	10.1	10.6	-0.5
ASP119	3.3	-1.0	2.3	3.2	-0.9	Arg39	11.7	0.4	12.1	13	-0.9
GLU7	4.1	-1.3	2.8	2.9	-0.1	Lys41	8.2	0.4	8.6	10.8	-2.2
GLU35	4.9	-0.7	4.2	6.2	-2.0	Arg42	11.2	0.5	11.7	13.4	-1.7
HSP15	5.6	-0.7	4.9	5.7	-0.8	Lys46	10.6	-0.1	10.5	10.6	-0.1
RMSD					0.7	Glu49	4.3	-0.8	3.5	3.6	-0.1
						Asp50	4.6	-1.1	3.5	3.0	0.5
Barnase						Arg53	11.3	1.1	12.4	13.9	-1.5
Asp8	2.3	-1.3	1.0	2.9	-1.9	RMSD					1.2
Asp12	4.5	-1.0	3.5	3.8	-0.3						
His18	8.1	-0.2	7.9	7.9	0.0						
Asp22	4.3	-0.7	3.6	3.3	0.3	**RNaseA**					
Glu29	1.6	-1.1	0.6	3.8	-3.2	Glu2	4.6	-1.3	3.2	2.8	0.4
Asp44	4.1	-0.6	3.5	3.4	0.1	Glu9	4.6	-0.5	4.0	4	0.0
Asp54	2.3	-1.4	0.9	3.31	-2.4	His12	8.0	-0.6	7.4	6.2	1.2
Glu60	4.8	-1.1	3.6	3.4	0.2	Asp14	3.6	-1.1	2.5	2	0.5
Glu73	3.4	-1.3	2.1	2.1	0.0	Asp38	4.2	-1.4	2.7	3.5	-0.8
Asp75	5.4	-1.3	4.1	3.1	1.0	His48	8.2	0.4	8.7	6	2.7
Asp86	2.5	-1.0	1.5	4.2	-2.7	Glu49	4.0	-0.5	3.5	4.7	-1.2
RMSD					1.6	Asp53	3.6	-0.7	2.9	3.9	-1.0
						Asp83	4.6	-1.1	3.5	3.5	0.0
Ovomucoid						Glu86	5.0	-0.9	4.1	4.1	0.0

Asp7	5.0	-0.4	4.6	2.7	1.9	His105	6.4	-0.3	6.1	6.7	-0.6
Glu10	5.4	-0.8	4.6	4.1	0.5	Glu111	3.7	-0.9	2.8	3.5	-0.7
Glu19	5.7	-1.1	4.7	3.2	1.5	His119	5.8	-0.7	5.2	6.1	-0.9
Asp27	4.9	-1.2	3.7	2.3	1.4	Asp121	3.2	-1.2	2.0	3.1	-1.1
Glu43	5.2	-0.3	5.0	4.8	0.2	RMSD					1.0
His52	5.9	-0.2	5.8	7.5	-1.7						
					1.3						
TRX-red						**TRX-ox**					
Glu6	4.1	-0.1	4.0	4.8	-0.8	Glu6	3.6	-0.3	3.3	4.9	-1.6
Glu13	4.2	-0.5	3.7	4.4	-0.7	Glu13	4.7	-0.7	3.9	4.4	-0.5
Asp16	4.9	-0.2	4.7	4	0.7	Asp16	4.1	-0.5	3.6	4.2	-0.7
Asp20	6.2	-1.2	5.0	3.8	1.2	Asp20	4.9	-1.3	3.6	3.8	-0.2
Asp26	9.3	0.6	9.9	9.9	0.0	Asp26	8.3	0.9	9.1	8.1	1.0
Glu47	4.3	-0.5	3.8	4.1	-0.3	Glu47	4.3	-0.9	3.5	4.3	-0.8
Glu56	5.1	-1.1	4.0	3.3	0.7	Glu56	5.2	-1.1	4.1	3.3	0.8
Asp58	2.4	-1.0	1.4	5.3	-3.9	Asp58	5.0	1.3	6.3	5.2	1.1
Asp60	5.2	0.7	5.9	2.8	3.1	Asp60	2.8	-0.9	1.9	2.7	-0.8
Asp61	3.7	0.0	3.7	4.2	-0.5	Asp61	3.2	-0.3	2.9	3.9	-1.0
Asp64	3.5	-0.9	2.6	3.2	-0.6	Asp64	4.5	-0.1	4.4	3.2	1.2
Glu68	5.6	-0.8	4.8	4.9	-0.1	Glu68	5.2	-1.1	4.1	5.1	-1.0
Glu70	4.4	-0.9	3.5	4.6	-1.1	Glu70	4.4	-1.1	3.3	4.8	-1.5
Glu88	5.6	-1.1	4.5	3.7	0.8	Glu88	6.5	-1.1	5.4	3.6	1.8
Glu95	4.4	-1.4	3.0	4.1	-1.1	Glu95	4.2	-0.9	3.3	4.1	-0.8
Glu98	5.3	-0.6	4.8	3.9	0.9	Glu98	5.5	-0.8	4.7	3.9	0.8
Glu103	4.7	-0.5	4.3	4.4	-0.1	Glu103	4.3	-0.6	3.7	4.5	-0.8
RMSD					1.4	RMSD					1.0

Table 10 All individual calculations for 87 residues of 7 proteins after 20 ns of MD simulations. 20 conformations for every 1 ns were used and the average pK_a were calculated.

Proteins	<$\Delta pK_{a,int}$>	<$\Delta pK_{a,app}$>	$\Delta pK_{a,app}{}^{X\text{-}ray,\varepsilon_P=4}$
Lysozyme	0.7	1.1	1.7
RNaseA	1.1	1.0	1.5
Ovomucoid	1.8	1.2	0.8
Barnase	2.0	1.3	3.6
Thioredoxin-Ox	1.1	1.0	1.2
Thioredoxin-red	1.4	1.4	2.4
BPTI	1.3	0.9	0.9
Total	1.3	1.2	2.0

Table 11 Summary of RMSD of ΔpK_a between experimental pK_a and calculated intrinsic and apparent pK_a at , ε_P= 4 by averaging them over trajectories from 10 ns of MD simulations. W_{ij} is the averages of absolute values of the shifts by charge-charge couplings. For comparison, the result without MD simulation sampling is also listed

After sampling multiple conformations, the overall accuracy of the prediction was significantly improved from the results with only single structures. The majority improvements were achieved in intrinsic pK_a calculation while the shifts by W_{ij} were in the similar range as the calculations with single structures. Many other groups have incorporated Monte-Carlo simulation into PB model to take account for protein flexibilities.[25, 26, 48] Here, we also see the improvement by incorporating MD simulation into the PB method. Since we decouple charge-charge interactions and do conformational sampling, we should be able to use a small ε_P which will compensate for only missing induced dipole interaction, quantum entities, or other small electrostatic effects that this model does not capture.

In detail, lysozyme, barnase, and reduced thioredoxin especially show much smaller deviations from the experimental data when we use MD conformational sampling. For example, in case of lysozyme without conformational sampling, even though there was a big improvement in pK_a prediction for ASP66 compared to the classic PB method, big deviations of 2.7 pK_a unit from the experimental pK_a was observed. This large errors was corrected to -1.3 unit error after the samplings. This residue is buried and surrounded by many hydroxyl group and a better desolvation effect could have been captured by MD simulations. Glu73 and ASP75 of barnase, which affected the overall accuracy significantly, also showed much improvement after the sampling. With the single X-ray structure, the deviations from experimental the pK_a were 8.8 for GLU73 and 6.6 for ASP75 which is a very undesirable result. GLU73 which is exposed to solvent was corrected and this may be explained by correct dielectric boundaries obtained by MD samplings . ASP75 is a buried residue and very close to the side chain of ARG83 within 2 Å. In X-ray structure calculation, the intrinsic pK_a showed extremely high shifts which suggests overestimation of the desolvation effect despite the presence of the arginine group nearby. The standard deviation of the intrinsic pK_a of this residue over the trajectories was 1.27 unit which is a larger fluctuation than most cases. Therefore, statistical sampling can resolve such errors that can appear in a static protein structure. These results indicate that the pK_a of ionizable groups both on the surface and in the buried sites can be more reliably evaluated by considering protein relaxation effect.

Another way to evaluate the validity of MD conformational sampling is by comparing pK_a calculations between two different X-ray structures for the same protein to see if the results converge to each other. We used 2 PDB structures for hen egg lysozyme, 1HEL and 2LZT. The comparison of deviations from the experimental pK_a is listed in Table 12. For both structure, better correlations with experimental data were obtained after MD sampling. Although the overall accuracy is similar to each other for single structures, opposite predictions for ASP66 were observed. The pK_a's of this group for both structures were predicted in the same direction after MD simulation. To see if the calculations converge regardless of the accuracy, the calculated pK_a for 27 ionizable residues, including arginine, lysine, and histidines whose experimental pK_a is not available, are plotted in Figure 10. It is clearly shown that the numbers were predicted in a more narrow range from each other with higher R^2 value than when calculated with single structures. Therefore, MD conformational sampling also gets rid of the variability of single original structures and enables one to get more robust predictions.

Now, the question remains if large pK_a shifts can be more accurately predicted with consideration of the protein relaxation effect. The predictions and comparisons with the results from single structure are listed in Table 13. Although the predictions for ASP 26 of both thioredoxin and LYS 66 of staphylococcal nuclease were improved, the predictions for two other histidine cases got worse. During the MD simulation, these two cases have been stabilized and the large pK_a shifts were underestimated with the averaged structures. The difficulty of predicting pK_a of histidines with MD conformational sampling is discussed in detail later. Despite the worsen predictions for these histidines, the similar total RMSD was obtained and it was still shown to be better than the results at ε_P= 10 with the classic PB method. As a result, our data show that our method is accurate in predicting both the large pK$_a$ shifts and other normal cases using low dielectric constants.

To verify that a low dielectric constant is more consistent when it takes account into the protein relaxation effect, in contrast to the classic PB model, we tested our model at ε_P= 10 as well. (Table 14, 16) As seen in Table 14, the values obtained at ε_P= 10 were accurate and similar to those obtained at ε_P= 4 which shows that our method can use a small dielectric constant and still reproduce experimental pK$_a$. Another important point is that there was much more improvement when looking at a comparison between single structures to averaged structures at ε_P= 4 than at ε_P= 10. This leads us to the question of null model again. Since we are already dealing with the protein relaxation effect, high dielectric constants would underestimate other missing electrostatic effects even more. This explanation is supported in the observed worsening for the predictions of large pK$_a$ shifts at ε_P= 10 with the MD conformational sampling as seen in Table 15. Not only did the lower ε_P predict these cases much more accurately, but the RMSD value observed at ε_P= 10 was significantly worse than the results obtained with single structures which has been shown in Table 8.

Therefore, we conclude that our method with MD conformational sampling can effectively and consistently predict pK_a at $\varepsilon_P = 4$ for both normal cases and large pK_a shifts without having to worry about which dielectric constants we would need to use.

		With conformation sampling		With single X-ray structure	
ID	RES	ΔpKa^{2lzt}	ΔpKa^{1hel}	ΔpKa^{2lzt}	ΔpKa^{1hel}
18	ASP	0.7	-0.7	-1.0	-0.8
48	ASP	-1.5	-2.0	2.2	1.0
52	ASP	1.0	-0.2	-1.0	0.2
66	ASP	-2.3	-1.3	-2.1	2.7
87	ASP	0.4	0.0	0.4	0.6
101	ASP	0.2	-1.1	0.0	-0.2
119	ASP	-1.0	-0.9	-0.9	-0.5
7	GLU	0.4	-0.1	0.6	2.4
35	GLU	-1.1	-2.0	-1.4	-1.1
15	HSE	-0.1	-0.8	0.9	1.0
RMSD		1.1	1.1	1.2	1.3

Table 12 pK_a calculation comparisons between 1HEL and 2LZT which are the same protein, hen egg lysozyme. Both single structure and MD simulation sampling were tested. Listed numbers are the deviations from experimental pK_a.

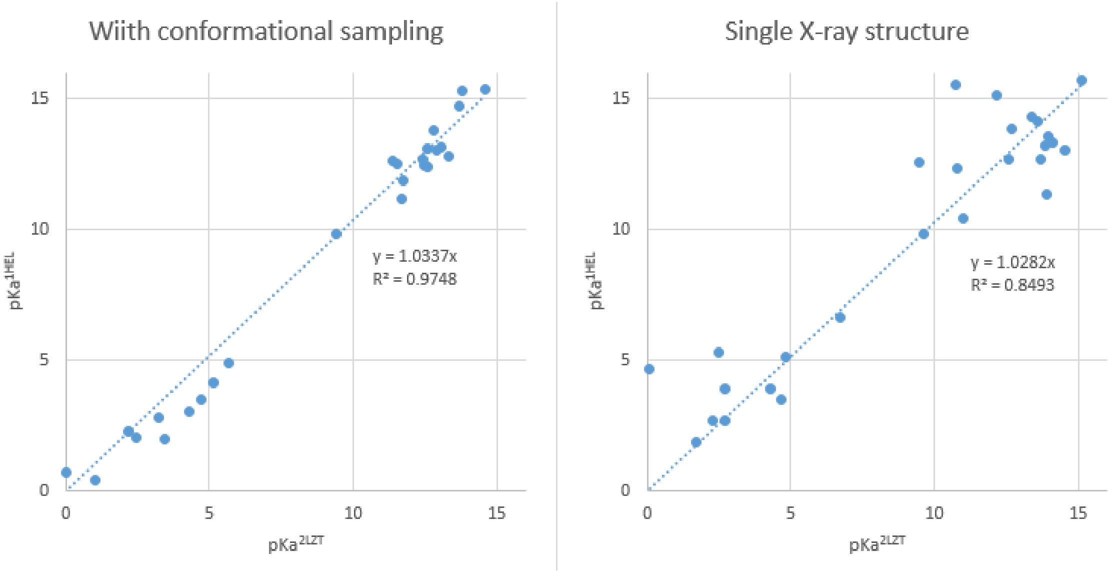


Figure 10 Scattered plots of the calculated pK_a of 2LZT vs 1HEL. Both single structure and MD simulation sampling were tested

				pKa^{calc}			$\Delta pKa^{calc\text{-}exp}$	
Proteins	Residue	pKa^{mod}	pKa^{exp}	pKa^{int}	Wij	pKa^{app}	$\Delta pK_{a,app}$	$\Delta pK_{a,app}{}^{Xray}$
Thioredoxin (red)	ASP 26	3.86	9.9	9.3	0.6	9.9	0.0	0.3
Thioredoxin (ox)	ASP 26	3.86	8.1	8.3	0.4	8.7	0.6	1.2
Staph. Nuclease	LYS 66	10.53	5.6	6.5	-0.4	6.1	0.5	-1.5
Erabutoxin	HIS 6	6.01	2.3	6.1	-2.0	4.2	2.0	-0.2
cytochrome c	HIS 26	6.01	2.6	4.4	-0.2	4.2	1.5	-0.3
						RMSD	1.1	0.9

Table 13 Our methods with MD conformational sampling for large pK_a shift cases

Proteins	<Apparen> ε = 4	<Apparent> ε = 10	Apparent, X-ray, ε = 4	Apparent, X-ray, ε = 10
Lysozyme	1.1	1.0	1.3	1.2
RNaseA	1.0	1.0	2.1	0.9
Ovomucoid	1.2	0.8	0.8	0.6
Barnase	1.3	1.0	1.5	1.6
Thioredoxin-Ox	1.0	1.0	2.5	1.2
Thioredoxin-red	1.4	1.4	0.9	1.7
BPTI	1.2	0.4	3.6	0.3
Total	1.2	1.0	1.9	1.2

Table 14 RMSD between experimental pK_a and the predicted pK_a with our method at two different dielectric constants

				pKa^{calc}		$\Delta pKa^{calc\text{-}exp}$	
Proteins	Residue	pKa^{mod}	pKa^{exp}	ε = 4	ε = 10	ε = 4	ε = 10
Thioredoxin (red)	ASP 26	3.86	9.9	9.9	5.4	0.0	-2.7
Thioredoxin (ox)	ASP 26	3.86	8.1	8.7	6.3	0.6	-3.6
Staph. Nuclease	LYS 66	10.53	5.6	6.1	8.5	0.5	2.9
Erabutoxin	HIS 6	6.01	2.3	4.2	4.8	2.0	2.5
cytochrome c	HIS 26	6.01	2.6	4.2	0.5	1.5	-2.1
					RMSD	1.1	3.3

Table 15 RMSD between experimental pK_a and the predicted pK_a with our method at two different dielectric constants

3.4 Comparison to other benchmarks

After we verified the validity of our method, we compared our results to other benchmarks. First, since our method was motivated by Sham *et al.* [23] which decoupled charge-charge interactions in PDLD/S model and showed a very good agreement with experimental data, we evaluate our W_{ij} comparing to the results in the previous work for lysozyme. (Table 15) No large perturbation by charge-charge interaction is observed in both predictions which are the desirable results as addressed in section 3.3. Our implementation has larger shifts for ASP52, ASP66, and ASP87. But for these cases, larger shifts help predict the experimental pK_a better.

Now, we compare our model to two other PB based benchmarks The first is H++ which is a webserver where one can quickly calculate the pK_a of a submitted protein. [18-20] As recommended by them as an optimal value, we used a single dielectric constant of ε_P= 10. Another benchmark has been reported in Nielsen *at al*.[48] This work incorporated Monte-Carlo simulation sampling for protein relaxation effects using DelPhi II. They used dielectric constant of 8 for most of the calculations and 16 for special criteria. The comparisons are listed in Table 15. We could reproduce the similar accuracy to theirs using lower dielectric constant. However, to our tests with the classic PB model, H++ failed to reproduce the large pK_a shifts.

residue	Our Wij	Sham *et al.*
GLU7	-1.3	-1.4
ASP18	-1.4	-0.9
GLU35	-0.7	-0.5
ASP48	-1.3	-1.0
ASP52	-0.7	-0.1
ASP66	-1.2	-0.6
ASP87	-1.1	-0.5
ASP101	-1.0	-1.3
ASP119	-1.0	-1.0
avg	1.1	0.8

Table 16 Comparison of calculated W_{ij} between ours and the results from Sham *at al* [23]

Proteins	Our method	H++	Nielsen et al.
Lysozyme	1.1(2.0)	1.0(1.6)	1.2(2.6)
RNaseA	1.0(2.7)	1.1(2.5)	1.0(2.4)
Ovomucoid	1.3(1.9)	0.7(1.0)	1.2(2.6)
Barnase	1.6(3.2)	1.4(3.1)	-
Thioredoxin-Ox	0.9(1.8)	1.0(4.2)	-
Thioredoxin-red	1.0(3.9)	1.0(4.6)	-
BPTI	0.9(2.2)	0.8(2.2)	0.7(2.0)
Total	1.2	1.4	1

Table 17 Comparison of out method to other benchmarks. Listed numbers are RMSD from experimental data and the largest errors are listed in bracket

3.5 Limitations and other challenges

While not the main focus of this study, we found several important discrepancies depending on the parameters set in the calculation. One is the dielectric boundary conditions. There are several ways to define the boundary between proteins and solvents. It can be defined by either the van der Waals surface or molecular surface. Theoretically, even though using the molecular surface which takes account for accessibility of solvents is more physically sound, we often observed better accuracy when using the van der Waals surface. This has been previously addressed elsewhere.[25, 49] We occasionally observed significantly large differences between two results with different boundary settings. In this study, we chose the conditions which generated the smaller perturbation for each protein.

Another difficulty we faced was the convergence problem in the calculation of W_{ij} which resulted in unacceptable huge pK_a shifts. It is likely that this was due to systemic errors caused by the sequential calculation for each residue. The Coulombic interaction energies were calculated in the order of the residue number as defined in the PDB file. As a result, the calculation can be trapped in a fluctuation between two numbers. This can be solved either by giving a different number of iterative steps to choose the smaller perturbation around 1 pK_a unit shift or by calculating the energies in different order of the residues. However, more consistent method needs to be devised to effectively remove this problem.

We wanted to stress the pK_a calculation of histidine. Many times, histidine should be treated in a special way since its side chain has two possible protonation sites(HSD and HSE) and it can have a flipped configuration. Especially, the calculated numbers can be very different between before and after MD conformational sampling. We usually calculate the pK_a with single X-ray structure first to start with the initial protonated states that have smaller pK_a shifts which are suggested to be more stable. Then, MD simulation is performed with these states. However, it sometimes turns out that it actually becomes destabilized during the MD simulation and we have to perform the simulation with the other protonated states. This can be very crucial in the studies of proteins where histidine plays a very important role in protein stability and conformational change such as Dengue virus envelope protein.[50, 51] Similarly, glutamic acid and aspartate acid have two possible protonation sites in the carboxyl groups. Even though the alternative protonation does not matter during MD simulation in this case since they are simulated in charged states, significantly different pK_a values are often observed in the calculation depending on which site is protonated. Thus, proper protonation site needs to be selected carefully.

Lastly, our implementation actually includes the calculations for serine and tyrosine but the results were not satisfying. One possible scenario is that CHARMM27 parameter defines the radius of protonated and deprotonated oxygen in their side chain in the same sizes. After trying different radius and partial charges, we found that the result is very sensitive to these parameter values. Also, more experimental data sets are required to evaluate the calculations for these two residues more reliably.

Conclusions

This study presents a more reliable and robust calculation with PB methods by decoupling charge-charge interactions and incorporating explicit

MD conformational sampling. Our method takes away the problem of adjusting dielectric constants which inevitably causes a loss of accuracy either in normal cases or large pK_a shifts cases. We tested our method that incorporated the PDLD/S approach against the classic PB model which has been initially suggested by Warshel and coworkers. There have been a lot of efforts to improve the pK_a prediction with PB model for decades by many other considerations such as optimizing hydrogen bond, other parameters, and, most importantly, trying to find an ‘optimal’ dielectric constant. Our work contributes to narrowing down these considerations by eliminating this dependence of dielectric constants in PB model.

Reference

1. Warshel, A. and R.M. Weiss, *Energetics of Heme-Protein Interactions in Hemoglobin.* Journal of the American Chemical Society, 1981. **103**(2): p. 446-451.
2. Warshel, A. and S.T. Russell, *Calculations of Electrostatic Interactions in Biological-Systems and in Solutions.* Quarterly Reviews of Biophysics, 1984. **17**(3): p. 283-422.
3. Matthew, J.B., *Electrostatic Effects in Proteins.* Biophysical Journal, 1985. **47**(2): p. A20-A20.
4. Nakamura, H., *Roles of electrostatic interaction in proteins.* Q Rev Biophys, 1996. **29**(1): p. 1-90.
5. Tanford, C. and J.G. Kirkwood, *Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres.* Journal of the American Chemical Society, 1957. **79**(20): p. 5333-5339.
6. Lee, F.S., Z.T. Chu, and A. Warshel, *Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs.* Journal of Computational Chemistry, 1993. **14**(2): p. 161-185.
7. Gilson, M.K. and B. Honig, *Calculation of the Total Electrostatic Energy of a Macromolecular System - Solvation Energies, Binding-Energies, and Conformational-Analysis.* Proteins-Structure Function and Genetics, 1988. **4**(1): p. 7-18.
8. Gilson, M.K., K.A. Sharp, and B.H. Honig, *Calculating the Electrostatic Potential of Molecules in Solution - Method and Error Assessment.* Journal of Computational Chemistry, 1988. **9**(4): p. 327-335.
9. Nicholls, A. and B. Honig, *A Rapid Finite-Difference Algorithm, Utilizing Successive over-Relaxation to Solve the Poisson-Boltzmann Equation.* Journal of Computational Chemistry, 1991. **12**(4): p. 435-445.
10. Bashford, D. and D.A. Case, *Generalized born models of macromolecular solvation effects.* Annual Review of Physical Chemistry, 2000. **51**: p. 129-152.
11. Onufriev, A., D. Bashford, and D.A. Case, *Modification of the generalized Born model suitable for macromolecules.* Journal of Physical Chemistry B, 2000. **104**(15): p. 3712-3720.
12. Rocchia, W., E. Alexov, and B. Honig, *Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions.* Journal of Physical Chemistry B, 2001. **105**(28): p. 6507-6514.
13. Rocchia, W., et al., *Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects.* Journal of Computational Chemistry, 2002. **23**(1): p. 128-137.
14. Brooks, B.R., et al., *CHARMM: The Biomolecular Simulation Program.* Journal of Computational Chemistry, 2009. **30**(10): p. 1545-1614.
15. Baker, N.A., et al., *Electrostatics of nanosystems: Application to microtubules and the ribosome.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(18): p. 10037-10041.
16. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Adv Protein Chem, 2003. **66**: p. 27-85.
17. Salomon-Ferrer, R., D.A. Case, and R.C. Walker, *An overview of the Amber biomolecular simulation package.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2013. **3**(2): p. 198-210.

18. Gordon, J.C., et al., *H++: a server for estimating pKas and adding missing hydrogens to macromolecules.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W368-71.
19. Myers, J., et al., *A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules.* Proteins-Structure Function and Bioinformatics, 2006. **63**(4): p. 928-938.
20. Anandakrishnan, R., B. Aguilar, and A.V. Onufriev, *H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W537-41.
21. Jo, S., et al., *CHARMM-GUI: a web-based graphical user interface for CHARMM.* J Comput Chem, 2008. **29**(11): p. 1859-65.
22. Simonson, T. and C.L. Brooks, *Charge Screening and the Dielectric Constant of Proteins: Insights from Molecular Dynamics.* Journal of the American Chemical Society, 1996. **118**(35): p. 8452-8458.
23. Sham, Y.Y., Z.T. Chu, and A. Warshel, *Consistent Calculations of pKa's of Ionizable Residues in Proteins: Semi-microscopic and Microscopic Approaches.* The Journal of Physical Chemistry B, 1997. **101**(22): p. 4458-4472.
24. Warshel, A. and A. Papazyan, *Electrostatic effects in macromolecules: fundamental concepts and practical modeling.* Current Opinion in Structural Biology, 1998. **8**(2): p. 211-217.
25. Alexov, E., et al., *Progress in the prediction of pKa values in proteins.* Proteins, 2011. **79**(12): p. 3260-75.
26. Teixeira, V.H., et al., *On the use of different dielectric constants for computing individual and pairwise terms in poisson-boltzmann studies of protein ionization equilibrium.* J Phys Chem B, 2005. **109**(30): p. 14691-706.
27. Antosiewicz, J., J.A. McCammon, and M.K. Gilson, *Prediction of pH-dependent properties of proteins.* J Mol Biol, 1994. **238**(3): p. 415-36.
28. Schutz, C.N. and A. Warshel, *What are the dielectric "constants" of proteins and how to validate electrostatic models?* Proteins, 2001. **44**(4): p. 400-17.
29. Sham, Y.Y. and A. Warshel, *The surface constraint all atom model provides size independent results in calculations of hydration free energies.* Journal of Chemical Physics, 1998. **109**(18): p. 7940-7944.
30. Bashford, D. and M. Karplus, *pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model.* Biochemistry, 1990. **29**(44): p. 10219-10225.
31. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
32. Bashford, D., *An object-oriented programming suite for electrostatic effects in biological molecules An experience report on the MEAD project*, in *Scientific Computing in Object-Oriented Parallel Environments*, Y. Ishikawa, et al., Editors. 1997, Springer Berlin Heidelberg. p. 233-240.
33. Cerutti, D.S., et al., *Staggered Mesh Ewald: An Extension of the Smooth Particle-Mesh Ewald Method Adding Great Versatility.* Journal of Chemical Theory and Computation, 2009. **5**(9): p. 2322-2338.
34. Ryckaert, J.-P., G. Ciccotti, and H. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.
35. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.

36. Webb, H., et al., *Remeasuring HEWL pK(a) values by NMR spectroscopy: methods, analysis, accuracy, and implications for theoretical pK(a) calculations.* Proteins, 2011. **79**(3): p. 685-702.

37. Baker, W.R. and A. Kintanar, *Characterization of the pH titration shifts of ribonuclease a by one- and two-dimensional nuclear magnetic resonance spectroscopy.* Archives of Biochemistry and Biophysics, 1996. **327**(1): p. 189-199.

38. Schaller, W. and A.D. Robertson, *pH, ionic strength, and temperature dependences of ionization equilibria for the carboxyl groups in turkey ovomucoid third domain.* Biochemistry, 1995. **34**(14): p. 4714-23.

39. Forsyth, W.R., et al., *Theoretical and experimental analysis of ionization equilibria in ovomucoid third domain.* Biochemistry, 1998. **37**(24): p. 8643-52.

40. Sali, D., M. Bycroft, and A.R. Fersht, *Stabilization of protein structure by interaction of alpha-helix dipole with a charged side chain.* Nature, 1988. **335**(6192): p. 740-3.

41. Oliveberg, M., V.L. Arcus, and A.R. Fersht, *pKA values of carboxyl groups in the native and denatured states of barnase: the pKA values of the denatured state are on average 0.4 units lower than those of model compounds.* Biochemistry, 1995. **34**(29): p. 9424-33.

42. Qin, J., G.M. Clore, and A.M. Gronenborn, *Ionization equilibria for side-chain carboxyl groups in oxidized and reduced human thioredoxin and in the complex with its target peptide from the transcription factor NF kappa B.* Biochemistry, 1996. **35**(1): p. 7-13.

43. Brown, L.R., et al., *The influence of a single salt bridge on static and dynamic features of the globular solution conformation of the basic pancreatic trypsin inhibitor. 1H and 13C nuclear-magnetic-resonance studies of the native and the transaminated inhibitor.* Eur J Biochem, 1978. **88**(1): p. 87-95.

44. Garcia-Moreno, B., et al., *Experimental measurement of the effective dielectric in the hydrophobic core of a protein.* Biophys Chem, 1997. **64**(1-3): p. 211-24.

45. Inagaki, F., et al., *Conformation of erabutoxins a and b in aqueous solution as studied by nuclear magnetic resonance and circular dichroism.* Eur J Biochem, 1978. **89**(2): p. 433-42.

46. Cohen, J.S., W.R. Fisher, and A.N. Schechter, *Spectroscopic studies on the conformation of cytochrome c and apocytochrome c.* J Biol Chem, 1974. **249**(4): p. 1113-8.

47. Russell, A.J., P.G. Thomas, and A.R. Fersht, *Electrostatic effects on modification of charged groups in the active site cleft of subtilisin by protein engineering.* J Mol Biol, 1987. **193**(4): p. 803-13.

48. Nielsen, J.E. and G. Vriend, *Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based pK(a) calculations.* Proteins, 2001. **43**(4): p. 403-12.

49. Tjong, H. and H.X. Zhou, *On the dielectric boundary in Poisson-Boltzmann calculations.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 507-514.

50. Mukhopadhyay, S., R.J. Kuhn, and M.G. Rossmann, *A structural perspective of the flavivirus life cycle.* Nat Rev Microbiol, 2005. **3**(1): p. 13-22.

51. Fritz, R., K. Stiasny, and F.X. Heinz, *Identification of specific histidines as pH sensors in flavivirus membrane fusion.* J Cell Biol, 2008. **183**(2): p. 353-61.

52. Lee, F.S., et al., *Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603.* Protein Eng, 1992. **5**(3): p. 215-28.

Appendix A

Linear Response Approximation

It is necessary to consider protein reorganization and relaxation during the charge process. Linear Response Approximation(LRA) has been introduced and adopted by many electrostatic computation approaches to achieve it efficiently. [6, 23, 52] In LRA, it is assumed that the curvatures of the free energy graphs of two different charged states (Figure 4) are identical. Because $\lambda_a = \lambda_b$ by this assumption, the following equation can be derived

$$\Delta G_{\mathrm{a}\rightarrow\mathrm{b}} = \frac{1}{2}\left[\langle V_{\mathrm{a}} - V_{\mathrm{b}}\rangle_{\mathrm{a}} + \langle V_{\mathrm{a}} - V_{\mathrm{b}}\rangle_{\mathrm{b}}\right] \qquad (1.6)$$

where ‘a’ and ‘b’ designates non-charged and charged sates, respectively and < > represensts average values from a set of conformations using statistical mechanics with Monte Carlo or Molecular Dynamic simulations. However, the free energy associated with the non-charged state is neglectible compared to the charged state. Therefore, Eq 1.6 is further simplified as

$$\Delta G_{\mathrm{a}\rightarrow\mathrm{b}} \cong \frac{1}{2}\langle \Delta V\rangle_{\mathrm{b}} \qquad (1.7)$$

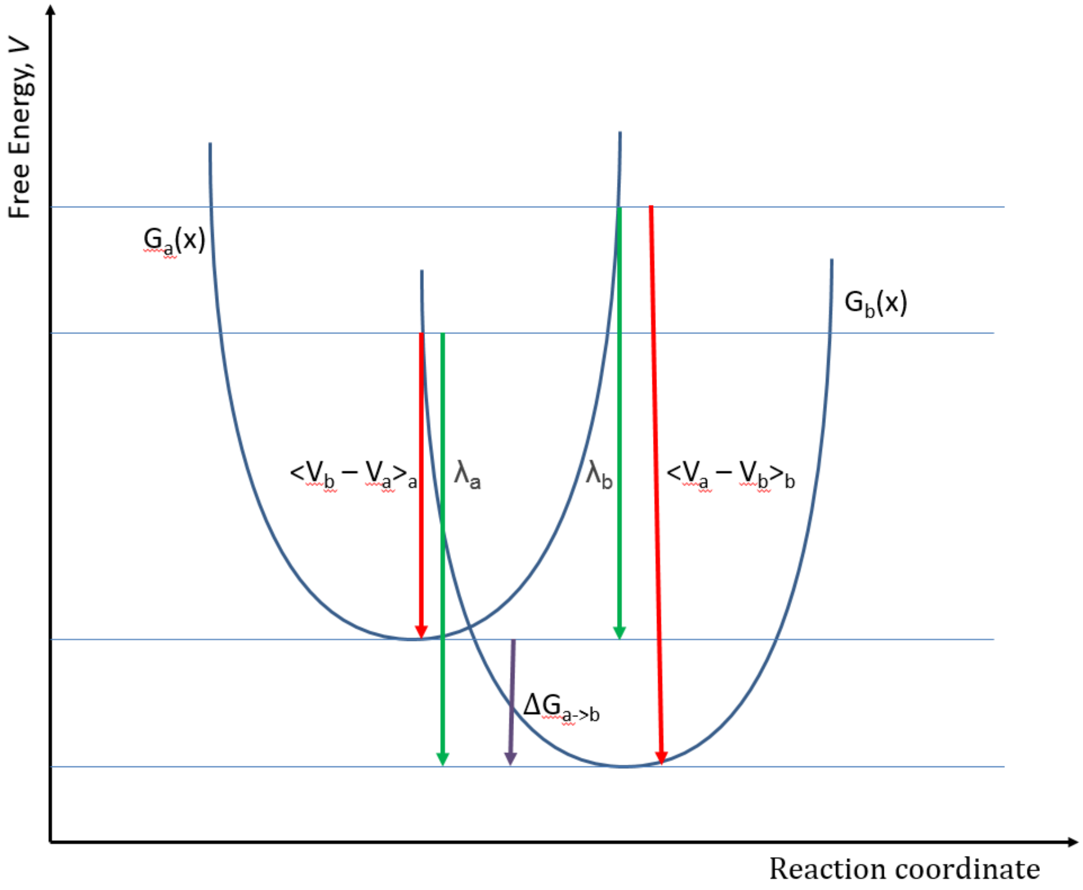


Figure A. Two parabolas describe the function of the free energy of the non-charged state(a) and the charged state(b). By the assumption of LRA, the curvatures of two free energy graphs are same. This leads to the linear relationship of $<V_b - V_a>_a + \Delta G_{a\rightarrow b}$ $=<V_a - V_b>_b - \Delta G_{a\rightarrow b}$ where $G_{a\rightarrow b}$ is the free energy associated with the adiabatic charging process