JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

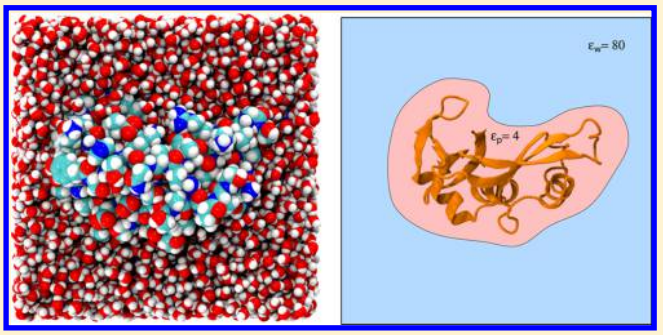
pK_a Values in Proteins Determined by Electrostatics Applied to Molecular Dynamics Trajectories

Tim Meyer and Ernst-Walter Knapp*

Institute of Chemistry and Biochemistry, Freie Universität Berlin, Fabeckstrasse 36A, 14195 Berlin, Germany

Supporting Information

ABSTRACT: For a benchmark set of 194 measured pK_a values in 13 proteins, electrostatic energy computations are performed in which pK_a values are computed by solving the Poisson−Boltzmann equation. In contrast to the previous approach of Karlsberg$^+$ (KB$^+$) that essentially used protein crystal structures with variations in their side chain conformations, the present approach (KB2$^+$MD) uses protein conformations from four molecular dynamics (MD) simulations of 10 ns each. These MD simulations are performed with different specific but fixed protonation patterns, selected to sample the conformational space for the different protonation patterns faithfully. The root-mean-square deviation between computed and measured pK_a values (pK_a RMSD) is shown to be reduced from 1.17 pH units using KB$^+$ to 0.96 pH units using KB2$^+$MD. The pK_a RMSD can be further reduced to 0.79 pH units, if each conformation is energy-minimized with a dielectric constant of ϵ_{min} = 4 prior to calculating the electrostatic energy. The electrostatic energy expressions upon which the computations are based have been reformulated such that they do not involve terms that mix protein and solvent environment contributions and no thermodynamic cycle is needed. As a consequence, conformations of the titratable residues can be treated independently in the protein and solvent environments. In addition, the energy terms used here avoid the so-called intrinsic pK_a and can therefore be interpreted without reference to arbitrary protonation states and conformations.



INTRODUCTION

Hydrogen bonds and protonation patterns of molecular systems depend on each other and are crucial to the structural integrity and function of proteins.[1−5] Therefore, determining the protonation pattern in proteins under different conditions is important. Since proteins are crowded with titratable groups that influence each other's protonation states, it is not an easy task to determine pK_a values in proteins experimentally. Moreover, it is virtually impossible to experimentally determine nonequilibrium protonation patterns involving transient excesses or deficiencies of protons occurring during enzymatic reactions or proton conduction.[6,7] Thus, computation of the energetics of protonation patterns in proteins can be particularly useful for providing insight in these areas. This has sparked an initiative among theoretical groups to foster the development and improvement of methods for computing pK_a values in proteins.[8] In the present work, this effort is continued by computing pK_a values in proteins by evaluating electrostatic energies of molecular systems using MD simulation data obtained with different protonation patterns.

The electrostatic approach is based on the essential assumption that the difference in pK_a values between the solvent and protein environments depends only on the electrostatic energies. This assumption works because the sum of van der Waals (vdW) interactions of titratable residues in protein and solvent environments is approximately the same and the intramolecular energies of titratable groups that are not of electrostatic origin are insensitive to the environment. The latter assumption is supported by the fact that *ab initio* pK_a computations of a large number of organic compounds yielded a pK_a root-mean-square deviation (pK_a RMSD) of only 0.5 pH units using the quantum chemical energies in vacuum combined with electrostatic energies of solvation for protonated and deprotonated molecular species.[9] This precision of computed pK_a values in water demonstrates that for organic compounds nonelectrostatic interactions are essentially environment-independent, i.e., they are approximately the same in solvent and protein environments. Even for the more challenging hexa-aqua metal complexes, the separation between quantum chemical energies in vacuum and electrostatic energies in solvent is approximately additive.[10] This is also due to the fact that the electrostatic energy terms necessary to compute pK_a values in proteins form double differences. These are the differences between protonated and deprotonated molecular species and between protein and solvent environments. As a consequence, many of the interaction terms cancel. Because the calculations are convenient and accurate, over the years many studies have used the electrostatic approach to compute pK_a values in proteins.[11−26]

Received: February 9, 2015
Published: May 5, 2015

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

The accuracy of computed pK_a values in proteins using electrostatic energy computations depends critically on the quality of high-resolution structures at appropriate pH values. Most of the protein structures available in the Protein Data Base[27] (PDB) have been solved at pH values around 7. If a structural change goes along with a protonation change at a titratable residue, then a single structure of a protein is not sufficient to determine a pK_a value reliably. Only in rare cases are structures of a protein available at different pH values, such as, for instance, the case of myoglobin.[28] If only one crystal structure is available, then the pK_a prediction procedure has to take care of this problem, e.g., by modeling structures corresponding to different pH values. In the recent past,[29] a major difficulty in computing accurate pK_a values for proteins was identified as the problem of breaking salt bridges when the pH drops below 4 or rises above 10. The software package Karlsberg$^+$ (KB$^+$)[29] has addressed this problem by introducing automatic modeling procedures that generate structures at different pH values, the so-called pH adapted conformations (PAC). With this method, the accuracy of pK_a computations has been shown to be significantly improved, reaching a pK_a-RMSD of about 1.1 pH units relative to measured pK_a values. However, the modeling protocol of KB$^+$ is still quite restrictive. Either the protein structure is changed only by optimizing the hydrogen bond network for pH 7 (we call this the standard protocol; see Materials and Methods) or else the side chain coordinates of salt-bridge residues are more extensively optimized for pH < 4 and pH > 10.

In the present study, the strategy of KB$^+$,[29] to automatically generate PACs and use them for electrostatic energy computations to obtain pK_a values is generalized using all-atom molecular dynamic (MD) simulation data with explicit water. Thus, the changes in the protein structure that can be addressed are not limited to a reorganization of the hydrogen-bond network and variations in salt-bridge geometries. Instead, any type of structural change that occurs during an MD simulation can be used. In the original KB$^+$,[29] method, each PAC is a single protein structure involving only local conformational changes of residues in salt-bridge geometries, but here it comprises an ensemble of several hundred structures that are frames of MD trajectories.

The idea to combine electrostatic energy computation with MD simulations for evaluating pK_a values is not new and has been discussed and applied extensively in the past in several variants.[30−41] We would like to particularly emphasize the constant pH MD simulation technique, where the protonation of titratable groups can change during an MD simulation run.[42−52] The electrostatic pK_a computations in proteins were reviewed recently.[53,24] More recently, Nilsson et al.[41] showed, with an approach similar to the one presented here, that structures from MD simulations can be used for electrostatic energy computations to predict pK_a values with very high accuracy. For the dimeric leucine zipper, they analyzed frames from three MD trajectories that differed in the protonation pattern of titratable residues. They chose the following three protonation patterns: (1) all titratable residues (Arg, His, Lys, Asp, Glu) charged, tyrosine neutral; (2) glutamic acid and tyrosine neutral, all others charged; (3) lysine and tyrosine neutral, all others charged. For each pH value, the protonation probability was averaged over all frames of an MD trajectory. These averaged results were then combined using a weighting function based on the protonation probabilities. In this work, we present a procedure that extends this concept and show results of an extensive benchmark computation involving 194 measured pK_a values of 13 proteins. Unlike the previous work described above,[41] the present study employs electrostatic energies to weight the data of different MD simulations in which the employed protonation pattern is based on preliminary pK_a values computed using the protein crystal structure. Furthermore, the current study investigates how variations in the procedure influence the accuracy of the computed pK_a values. These are (i) post processing of the MD structures by energy-minimizing them with different dielectric constants and (ii) using different sets of atomic radii. We also compare the accuracy of the pK_a value computations in the present study with the results obtained by the empirical prediction scheme PropKa.[54,55]

The reformulation of the electrostatic energy terms in this work avoids the use of a reference protonation state (or reference conformation state), as is necessary for the traditionally introduced intrinsic pK_a value.[11,22,29] This pK_a value relates to the initial simplification that only one titratable group is present in the molecular system, which requires the introduction of a reference protonation state and correction terms in the residue-pair interactions to generalize the electrostatic energy function to arbitrary protonation states. As a consequence, the interactions between titratable residues were not directly interpretable within the former approach, which has changed now. The present formulation needs no thermodynamic cycle and avoids unphysical electrostatic interactions between atom pairs that are connected by less than four covalent bonds. These unphysical interactions are falsely included in electrostatic energy computations using numerical solutions of the linearized Poisson−Boltzmann equation (*l*PBE). Their influence is critical if several protein conformations are used, as is done in the present work.

MATERIALS AND METHODS

Benchmark Set of Measured pK_a Values in Proteins. The procedure for calculating pK_a values was tested on a set of 13 proteins with known crystal structures and a sufficiently large number of measured pK_a values, 194 pK_a values in all. Eleven protein structures (PDB IDs: 4PTI, 1PGA, 1A2P, 2LZT, 3RN3, 2RN2, 1HNG, 3ICB, 1PPF, 1ERT, 1XNB) have been chosen from the set of 15 proteins used to demonstrate the performance of the program KB$^+$.[29] Since the procedure in the present study requires significantly more CPU time compared to that for KB$^+$, proteins from the prior study with less than seven measured pK_a values were not included here. To enlarge the benchmark set, we added two proteins: a leucine zipper (PDB ID 2ZTA[56]), since it was considered in related work computing pK_a values from MD simulation data,[41] and a staphylococcal nuclease (SNase) variant Δ+PHS (PDB ID 3BDC[57]), since it has been a central protein in recent efforts to compute pK_a values.[8] The majority of the measured pK_a values used in the current study were collected by Georgescu et al.[58] from the literature. For xylanase, the measured pK_a values are taken from Joshi et al.,[59] for the leucine zipper, they are provided by Matousek et al.,[60] and for the SNase variant Δ+PHS, by Castaneda et al.[57]

Necessary Modeling Steps for the Benchmark Set. All modeling steps were performed with the CHARMM[61] program using the CHARMM22 force field,[62] unless otherwise indicated. Ligands and ions in the crystal structures of the proteins were removed. Crystal water molecules were kept for the MD simulations, whereas they were removed for all

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

electrostatic energy computations, if not otherwise stated. In the case of the third domain of the turkey ovomucoid inhibitor (OMTKY3, PDB ID 1PPF), only chain I was used since the complete crystal structure also contains the protein PMN elastase. For staphylococcal nuclease variant Δ+PHS, the first six and last eight residues are disordered in the crystal structure. These residues were not added by modeling. Here, the termini of the crystal structure were neutralized with an acetylated N-terminus and a methylated C-terminus to avoid the introduction of artifactual charges on the residues at these positions. For the GCN4 leucine zipper, one Gly was added to the N-terminus by modeling and a Glu was added to the C-terminus to reproduce the protein used to measure pK_a values.[60] The crystal structure for xylanase (PDB ID 1XNB) was modified slightly, interchanging the coordinates of the ND2 and OD1 atoms of residue Asn35, which has a large influence on the pK_a value of the neighbor residue Glu172, as discussed later. For the protein rat T-lymphocyte adhesion glycoprotein CD2 (PDB id 1HNG), only part of the crystal structure involving residues 1−99 of chain A was used, since that part was also used to measure the pK_a values. In the case of thioredoxin (PDB ID 1ERT), the deeply buried crystal water 136 in chain A that is in contact with Asp26 was kept throughout all of the calculations, and all side chains of the protein were energy-minimized with CHARMM,[61] as was done in the previous benchmark computation[29] to compensate for the influence of crystal contacts. The minimization was done in two steps, first in vacuum and then with an implicit water model using the GBSW[63] module in CHARMM.[61]

pK_a Computations with Karlsberg$^+$. For comparison, all pK_a values in the benchmark set were also computed using the standard protocol of the software and web application KB$^+$.[29] This protocol consists in the automatic creation of 11 PACs for pH values of −10, 7, and 20. For all PACs, the structures are initially prepared by energy minimizing the oxygen atoms of all carboxylic groups. The essence of the procedure used by KB$^+$ is iterating between structure relaxation (for a given pH-dependent protonation pattern) and evaluation of the protonation pattern (for a given structure) until self-consistency is reached. Coordinates of hydrogen atoms are always energy-minimized using CHARMM,[61] and additionally for pH −10 and 20, the dihedrals of the side chains of the residues involved in salt bridges are randomly sampled and consecutively energy-minimized. The results obtained here with the KB$^+$ web application may differ slightly from the values published earlier[29] due to minor adjustments in the protocol and the random search for salt-bridge geometries. Besides the standard protocol of KB$^+$,[29] a simplified protocol is also available in KB$^+$, which is called the sc_{pH7} procedure. It uses the protein crystal structure with all waters removed and optimizes only the hydrogen atom positions at pH 7 self-consistently.

Protocol of pK_a Computations. *General Overview.* The new protocol for calculating pK_a values introduced in the present study consists of five steps. In short, they comprise (1) manipulating and completing the protein structures as described above in the Necessary Modeling Steps for the Benchmark Set section, (2) selecting a pool of protonation patterns, (3) preparing and running MD simulations for each selected protonation pattern, (4) performing the electrostatic energy computations after optionally energy minimizing the structure of each frame of the MD trajectories, and (5) averaging the results of the electrostatic energy computations for each MD trajectory and combining the averaged results to obtain titration curves and pK_a values. Modeling and solvation with explicit water were performed with CHARMM.[61] Energy minimization and MD simulations were performed with NAMD,[64] version 2.9. Steps 2−5 are as follows.

Step 2: Protonation Pattern for MD Simulation. To prepare the MD simulation of a protein, for each titratable residue a complete protonation pattern has to be chosen. Such a set of protonation states is called a fixed protonation pattern, abbreviated here as FPP. Six residues are considered to be titratable, Asp, Glu, Lys, His, Cys, and Tyr, as well as the C- and N-termini (C-ter and N-ter). Since the residue Arg was not explicitly observed to be deprotonated in an intact native protein structure, we do not include it as titratable residue in the present study. In the electrostatic energy computations using the structures from these MD simulations, we deviate from the fixed protonation states and explore all possible protonation patterns.

The four selected FPPs, listed in Table 1, are the result of an extensive exploration of different possibilities. In the FPP denoted pH < 4, all residues are protonated (prot in Table 1), corresponding to a low pH value. Hence, in the FPP at pH < 4, Asp and Glu are charge neutral, His is positively charged, and Lys, Cys, and Tyr are protonated. The FPP denoted pH 5 is obtained by determining the protonation of the individual titratable residues (pK_a in Table 1) with the simplified protocol of KB$^+$, except for Lys, which is kept protonated (prot). The FPP at pH 7 is the same as that at pH 5, except that the acidic residues of Asp and Glu are kept deprotonated (deprot). For titratable residues that had been added by modeling procedures at the N- or C-terminus of a protein, the protonation state was determined using the pK_a values of these residues in aqueous solution in comparison with the relevant pH value. In the fourth FPP, denoted pH > 10, all titratable residues are kept deprotonated (deprot). Hence, in this case, His and Lys are charge neutral, whereas Cys and Tyr are negatively charged. As mentioned before, Arg was kept protonated throughout the present study; to treat it also as a titratable residue, one would require an MD simulation with a fifth FPP, where Arg is also deprotonated.

Table 1. Fixed Protonation Pattern (FPP) Used for the Eight Different Types of Titratable Residues in the Four MD Simulation Runs[a]

residue	pH < 4	pH 5	pH 7	pH > 10
Asp	prot	pK_a	deprot	deprot
Glu	prot	pK_a	deprot	deprot
Lys	prot	prot	pK_a	deprot
His	prot	pK_a	pK_a	deprot
Tyr	prot	pK_a	pK_a	deprot
Cys	prot	pK_a	pK_a	deprot
N-ter	prot	pK_a	pK_a	deprot
C-ter	prot	pK_a	pK_a	deprot

[a]Arginines are always kept protonated. N-ter and C-ter are the N- and C-termini of the protein, respectively. Residues that are protonated or deprotonated are labeled prot and deprot, respectively. The label pK_a indicates that the protonation is set according to the results of a pK_a computation with KB$^+$ based on the crystal structure.

Step 3: Preparing MD Simulation Runs. For each FPP defined in the previous step, an MD simulation is performed. To prepare these MD runs, the corresponding FPP is modeled and the protein is packed in a box of TIP3 water molecules with

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

periodic boundary conditions. The edge length of the cubic water box is equal to the maximum extension of the protein plus an additional 15 Å on all sides. First, only the water molecules and the hydrogen atoms of the protein were energy-minimized while the remaining atoms, whose positions are known from the crystal structure, were kept fixed. Then, a short MD simulation of 25 ps at 300 K without an explicit heating phase was performed with the same restraints to equilibrate the water molecules and hydrogen atoms before the non-hydrogen atoms of the protein were allowed to move. The temperature was controlled using Langevin dynamics with a friction constant of 1 ps^{-1}. After these preparations, all atoms were energy-minimized, and the system was heated by velocity rescaling to 300 K within 50 ps in steps of 20 K. Finally, the main MD production run of 10 ns was performed under NPT conditions using the Nosé–Hoover thermostat.[65] The first nanosecond of each MD run was used for equilibration and was not considered in the analysis. Subsequently, trajectory frames at 100 ps intervals were used for electrostatic energy computations, resulting in 90 structures per MD simulation. As a representative example, for lysozyme, the RMSDs of the protein backbone atoms relative to the crystal structure, averaged over the 10 ns, were 1.3, 1.1, 1.1, and 1.2 Å for the four MD simulations pH < 4, pH 5, pH 7, and pH > 10, respectively. The detailed time course of the RMSD values is shown in Figure 2.

Step 4: Performing the Electrostatic Energy Computations. As an optional step, each structure for the complete system (water and protein) sampled from the simulations was energy-minimized with 1000 steps of conjugate gradient combined with line search minimization, using the standard energy-minimization algorithm of NAMD 2.9,[64] before the electrostatic energy computations were performed. The energy minimization was performed with periodic boundary conditions in a homogeneous dielectric of either $\varepsilon = 1$ or $\varepsilon = 4$. These two variants are named $\varepsilon_{min} = 1$ and $\varepsilon_{min} = 4$ in the Results section. For each structure at time t_{frame}, electrostatic energy computations were performed using the software TAPBS,[29] a modified version of APBS,[66] which is part of KB^+,[29] as well as the software APBS itself. A detailed description how the electrostatic energies are evaluated follows in the next section.

Step 5: Averaging the Results of Electrostatic Energy Computations. For each MD simulation (MD_m), the electrostatic energy terms, introduced in the next section, are averaged over all sampled structures. These averaged electrostatic energy terms are used to compute titration curves with the software Karlsberg,[28] which applies a Monte Carlo (MC) algorithm to determine the energetically most favorable protonation states for each pH. From the titration curves, the pK_a value of each residue is determined as the pH where the titratable residue is protonated with probability 0.5.

Electrostatic Energy Computations. *General Considerations.* To obtain absolute pK_a values of titratable residues in the protein environment, the double difference of electrostatic free energies $\Delta\Delta G = \Delta G^h - \Delta G^d$ must be added to the measured pK_a^{exp} value according to[22,29]

$$pK_a^{protein} = pK_a^{exp} - \log_{10}(e)/RT \times \Delta\Delta G \quad (1)$$

where $\Delta G^h = G^h(\text{protein}) - G^h(\text{solvent})$ and $\Delta G^d = G^d(\text{protein}) - G^d(\text{solvent})$ are the free energy differences for transferring the protonated (h) or deprotonated (d) species, respectively, from the solvent to the protein environment. The measured pK_a^{exp} values and the nature of these compounds in aqueous solution are tabulated in ref 22. Such a simple expression is applicable if one considers an isolated titratable group in a molecular system involving a single conformation. The more precise expression that is derived below also considers the mutual interaction of different titratable groups and variations in protein conformation.

For the electrostatic energy computations, the protein is described as a dielectric continuum with $\varepsilon_{protein} = 4$ inside the protein volume and $\varepsilon_{solv} = 80$ outside, where $\varepsilon_{protein} > 1$ accounts for the dielectric shielding due to fine-grained structural variations not contained in the frames of MD simulation and polarization effects not explicitly captured by the fixed-charge force field. The protein volume is defined as the joint volume contained within the solvent excluded surface[67] (SES) of the individual protein atoms, which is generated by rolling a sphere of radius 1.4 Å over the vdW surface of the protein. An analogous procedure is applied to generate the boundary surface of a titratable residue (solute) in aqueous solution. Here, we also use $\varepsilon_{solute} = 4$ inside the solute and $\varepsilon_{solv} = 80$ outside. In both the protein and solute, the charges are represented by fixed atomic point charges, which are taken from the CHARMM force field.[62] The electrostatic potential generated by the charge cloud $\rho(\vec{r})$ (here represented by a set of atomic point charges) in an inhomogeneous dielectric medium with a spatially dependent dielectric constant $\varepsilon(\vec{r})$ and ionic strength $I(\vec{r})$ measured by the inverse Debye length $\kappa(\vec{r})$

$$\kappa^2(\vec{r}) = 8\pi e_0^2 I(\vec{r})/(RT) \quad (2)$$

is obtained by solving the linearized Poisson–Boltzmann equation (lPBE)

$$\vec{\nabla}[\varepsilon(\vec{r})\ \vec{\nabla}\ \Phi(\vec{r})] = -\ 4\pi\rho(\vec{r}) + \kappa^2(\vec{r})\ \Phi(\vec{r}) \quad (3)$$

A nonvanishing ionic strength for the protein in the solvent of $I = 0.1$ M is used for our computations. The Stern layer separating the ions in the solvent from the protein volume is generated by a rolling sphere with radius of 2 Å along the vdW surface. The electrostatic energy of a set of atomic partial charges $\{q(n)|n = 1, 2, \ldots N_q\}$ at positions in such an electrostatic potential is computed as

$$E(q,\rho) = \sum_{n=1}^{N_q} q(n)\ \Phi(\vec{r}_n) \quad (4)$$

Artifacts Occurring while Solving the lPBE Numerically. The numerical solution of the lPBE for molecular systems generates two types of artifacts. The first is the commonly known grid artifact, which arises since the atomic point charges are distributed among the nearest grid points. The factional charges on neighbor grid points belonging to the same atom start to interact. This interaction is artificial and increases with grid resolution. However, even if the point charge of an atom is precisely on a grid point, the electrostatic potential should be infinity at this grid point. In numerical computations solving the lPBE, the singularity is replaced by a very large value, which depends on grid geometry and resolution. This artifact is usually corrected by using an appropriate thermodynamic cycle.

The second artifact is the inclusion of the 1–2, 1–3, 1–4 atom-pair interactions of atoms that are separated by less than four covalent bonds in a molecule. In classical MD force fields, these unphysical electrostatic interactions are routinely

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827–2840

excluded as nonbonded exclusion. Due to the closeness of the participating atoms, the unphysical 1−2, 1−3, 1−4 electrostatic atom-pair interactions would contribute considerably to the total electrostatic energy. Thus, even small variations in atom-pair distances going along with changes in molecular conformation will have a marked influence on electrostatic energies. The latter artifact is routinely removed by a thermodynamic cycle procedure if only a single molecular conformation is considered. Below, we show how both artifacts are removed.

Characterizing Protonation States. The protonation state of a titratable residue can often not simply be described as being protonated (h) or deprotonated (d). For instance, histidine has two different deprotonated states where the remaining proton is bound to one of the two imidazole nitrogens. Glu and Asp have two different protonated states where the hydrogen is bound to one of the two acidic oxygens. We call these protonation-dependent tautomeric states protonation microstates (μ-states). In this study, we consider only the above-mentioned μ-states for His, Glu, and Asp. However, the present procedure allows additional μ-states to be introduced if needed. The μ-states of a titratable residue i are denoted $\alpha_k(i)$, where k gives the particular μ-state for each given protonation state $\alpha \in \{h,d\}$. To evaluate the electrostatic free energy $\Delta\Delta G$, eq 1, the following four types of electrostatic energy terms are needed.

(1). The Background Energy. $G^{\text{protein}}_{i,\text{back}}$ is the electrostatic free energy of residue i in the protein dielectric environment ($\varepsilon_{\text{protein}}$ = 4 inside and $\varepsilon_{\text{solv}}$ = 80 outside) interacting with all background charges of the protein, which are all of the charges in the protein excluding the atomic charges of all titratable residues. The background energies $\Delta G_{i,\text{back}}[\alpha_k(i),\text{MD}_m(t_{\text{frame}})]$ depend not only on the residue i but also on the adopted μ-state [$\alpha_k(i) = h_k(i)$ for protonated and $\alpha_k(i) = d_k(i)$, for deprotonated residue i]. $\text{MD}_m(t_{\text{frame}})$ indicates the dependence on trajectory frame t_{frame} of MD simulation m. All electrostatic energy terms possess this dependence, which in the following will not explicitly be indicated in the equations.

The background electrostatic energy of residue i in the protein environment is evaluated as

$$G^{\text{protein}}_{i,\text{back}}[\alpha_k(i)] = \sum_{n=1}^{N_{\text{back}}} q_{\text{back}}(n)\{\Phi^{\text{protein}}_{i,\alpha_k(i)}[\text{PB};\vec{r}(n)] - \Phi^{\text{homo}}_{i,\alpha_k(i)}[\text{PB};\vec{r}(n)]\} + G^{\text{homo}}_{i,\text{back}}[\text{Coulomb};\alpha_k(i)] \quad (5)$$

The first term in the sum involves also the unphysical 1−2, 1−3, 1−4 interactions. Since the second term removes all atom-pair interactions in the homogeneous dielectric, the unphysical interactions are also removed, and only the interactions of the charges with the reaction field at the protein−solvent interface remain. The last term using Coulomb's law places the atom-pair interactions back but leaves out the unphysical interactions by applying the nonbonded exclusion. $\Phi^{\text{protein}}_{i,\alpha_k(i)}[\text{PB};\vec{r}(n)]$ is the electrostatic potential at the positions $\vec{r}(n)$ of the N_{back} background charges $q_{\text{back}}(n)$, which is generated by the charges of residue i in μ-state α_k. It is evaluated by solving the *l*PBE, eq 3, in the protein environment ($\varepsilon_{\text{protein}}$ = 4 and $\varepsilon_{\text{solv}}$ = 80). $\Phi^{\text{homo}}_{i,\alpha_k(i)}[\text{PB};\vec{r}(n)]$ is the corresponding electrostatic potential for the homogeneous dielectric medium ($\varepsilon_{\text{protein}}$ = 4 = $\varepsilon_{\text{solv}}$). $G^{\text{homo}}_{i,\text{back}}[\text{Coulomb};\alpha_k(i)]$ is the electrostatic interaction between the charges of residue i in the μ-state $\alpha_k\{q_{i,\text{res}}[\alpha_k,n_i]\}$ and the background charges $\{q_{\text{back}}(n)\}$ evaluated using Coulomb's law for charge pair interactions in the homogeneous dielectric with ε = 4 everywhere, where the unphysical 1−2, 1−3, 1−4 atom-pair interactions are avoided, corresponding to the nonbonded exclusion list used in molecular force fields. These unphysical electrostatic interactions occur for atom pairs at the interface between protein backbone and residue side chain atoms. In the present application, we have for technical reasons only avoided the most dominant 1−2 atom-pair interaction.

(2). The Born Energy. This is the self-energy of a titratable residue located in solvent or protein environment, yielding for the μ-state α_k of residue i $G^{\text{solvent}}_{i,\text{Born}}[\alpha_k(i)]$ and $G^{\text{protein}}_{i,\text{Born}}[\alpha_k(i)]$, respectively. It is computed as follows

$$G^{\text{environment}}_{i,\text{Born}}[\alpha_k(i)] = G^{\text{environment}}_{i,\text{Born}}[\text{PB};\alpha_k(i)] - G^{\text{homo4}}_{i,\text{Born}}[\text{PB};\alpha_k(i)] + G^{\text{homo4}}_{i,\text{Born}}[\text{Coulomb};\alpha_k(i)] \quad (6)$$

Here, the superscript environment refers to either protein or solvent dielectric environment, whereas homo4 refers to the homogeneous dielectric medium with ε = 4 everywhere. The last term, labeled Coulomb, is evaluated using Coulomb's law for the atom-pair interactions in the homogeneous dielectric medium with ε = 4, applying the nonbonded exclusion. The last two terms in eq 6 are added to cancel the grid artifact occurring in evaluating the first term by solving the *l*PBE and to cancel the nonbonded exclusion terms, which are included in the first term. The first two terms labeled PB are evaluated by solving the *l*PBE and evaluating the sum

$$G^{\text{environment}}_{i,\text{Born}}[\alpha_k(i)] = \frac{1}{2}\sum_{n_i} q_{i,\text{res}}[\alpha_k(i),n_i]\,\Phi^{\text{environment}}_{i,\alpha_k(i)}[\vec{r}(n_i)] \quad (7)$$

where n_i runs over the atoms of residue i. Here, the superscript environment denotes either protein and solvent but also homo4 (i.e., homogeneous dielectric with ε = 4) as used in eq 6. $q_{i,\text{res}}[\alpha_k,n_i]$ is the atomic partial charges of residue i in the μ-state α_k. $\Phi^{\text{protein}}_{i,\alpha_k(i)}$ is the same electrostatic potential as the one used for the background electrostatic energies, eq 5. However, here the electrostatic potential $\Phi_i^{\text{environment}}$, eq 7, generated by the charges of residue i is used at the atom positions of the same residue. These are self-energies, which require using the factor $^1/_2$ to avoid double counting.

(3). The Residue-Pair Interaction. $W_{i,j}[\alpha_k(i),\beta_\lambda(j)]$ describes the electrostatic energy between two titratable residues i and j in their respective μ-states α_k and β_λ located in the dielectric environment of the protein. It is evaluated based on the same electrostatic potentials that are also used for the background energies, eq 5, as follows

$$W_{i,j}[\alpha_k(i),\beta_l(j)] = \sum_{n_i} q_{i,\text{res}}[\alpha_k(i),n_i]\,\Phi^{\text{protein}}_{j,\beta_l(j)}[\vec{r}(n_i)] \quad (8)$$

In practical applications, a symmetrized expression, which is the arithmetic mean of $W_{i,j}[\alpha_k(i),\beta_\lambda(j)]$ and $W_{j,i}[\beta_\lambda(j),\alpha_k(i)]$, is used to reduce numerical errors.

(4). The Conformational Energy. This accounts for changes in electrostatic interaction between the background charges in the protein, which occur with a change in protein conformation. For a given conformation, this energy is

$$G^{\text{protein}}_{\text{conf}} = G^{\text{protein}}_{\text{conf}}(\text{PB}) - G^{\text{homo4}}_{\text{conf}}(\text{PB}) + G^{\text{homo4}}_{\text{conf}}(\text{Coulomb}) \quad (9)$$

with

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

$$G_{\mathrm{conf}}^{\mathrm{environment}}(\mathrm{PB}) = \frac{1}{2}\sum_{n=1}^{N_{\mathrm{back}}} q_{\mathrm{back}}(n)\,\Phi_{\mathrm{back}}^{\mathrm{environment}}[\vec{r}(n)] \tag{10}$$

where N_{back} here is the total number of atomic charges in the protein not in a titratable residue. The superscript environment and the arguments PB and Coulomb in the above equation have the same meaning as that described in connection with eq 6. In contrast to earlier theory, the above expression of conformational energy formally uses the conformational reference state where all charges are moved to infinity, yielding a vanishing energy, as is commonly used in physical descriptions of electrostatic energies.

Like the Born energy term, eq 7, the conformational energy, eq 10, is a self-energy, which requires the factor $^1/_2$ to avoid double counting. The superscript environment in eq 10 refers to the inhomogeneous dielectric environment of the protein or the homogeneous dielectric (homo4) with $\varepsilon = 4$ everywhere. In eq 9, the first two terms labeled PB are evaluated by solving the *l*PBE, eq 3, to obtain the electrostatic potentials generated by all atomic partial charges in the protein except for the titratable residues using protein dielectric environment and homogeneous dielectric ($\varepsilon = 4$), yielding $\Phi_{\mathrm{back}}^{\mathrm{protein}}$ and $\Phi_{\mathrm{back}}^{\mathrm{homo4}}$, respectively. The last term in eq 9 is evaluated by using Coulomb's law in the same way as in the last term in eq 6.

As in the case of the Born energy, eq 6, the last two terms in eq 9 are needed to eliminate the grid artifact and the nonbonded exclusion terms. In contrast to earlier approaches,[29] there is no explicit dependence on the protonation states of the titratable residues in the conformational energy term, but the conformations, themselves, will depend on the protonation pattern used for the MD simulation. Pairwise interactions of titratable residues are included in the residue-pair interactions $W_{i,j}$, eq 8, whereas interactions of the titratable residues with the background charges of the protein are considered by the background energy term, eq 5.

Combining Protein and Solvent-Related Energy Terms. The background and Born energy of a residue i in the protein can be combined into an effective one-residue interaction term that depends on protonation and μ-state of the residue

$$G_{i,\mathrm{one}}^{\mathrm{protein}}[\alpha_k(i)] = G_{i,\mathrm{back}}^{\mathrm{protein}}[\alpha_k(i)] + G_{i,\mathrm{Born}}^{\mathrm{protein}}[\alpha_k(i)] \tag{11}$$

using the energy terms defined by the eqs 5 and 6, respectively.

The difference in the solvent Born energy of residue i between the protonated and deprotonated states averaged over the corresponding solvent conformations σ

$$\langle \Delta G_{\mathrm{Born},h-d}^{\mathrm{solvent}}(i)\rangle_\sigma = \langle G_{i,\mathrm{Born}}^{\mathrm{solvent}}[h(i)]\rangle_\sigma - \langle G_{i,\mathrm{Born}}^{\mathrm{solvent}}[d(i)]\rangle_\sigma \tag{12}$$

is the electrostatic contribution to the free energy of solvation. The averaging $\langle\rangle_\sigma$ can be performed using conformations from MD simulation in explicit water or the average energies may be parametrized for each type of titratable residue. In the present application, we use for the solvent the same residue conformations as for the protein. The experimental free energy difference between protonated and deprotonated states $(h - d)$ of the titratable residue i in aqueous solution at a given pH value is given as

$$\Delta G_{\mathrm{exp},h-d}^{\mathrm{solvent}}(i, \mathrm{pH}) = RT\ln(10)[\mathrm{pH} - \mathrm{p}K_{\mathrm{a}}^{\mathrm{exp}}(i)] \tag{13}$$

where T is the absolute temperature and R is the ideal gas constant. The nonelectrostatic free energy contribution of the difference between protonated and deprotonated states $(h - d)$ of titratable residue i in aqueous solution is given as

$$\Delta G_{\mathrm{nonelectrostatic},h-d}^{\mathrm{solvent}}(i, \mathrm{pH}) = \Delta G_{\mathrm{exp},h-d}^{\mathrm{solvent}}(i, \mathrm{pH}) - \langle \Delta G_{\mathrm{Born},h-d}^{\mathrm{solvent}}(i)\rangle_\sigma \tag{14}$$

The general assumption is that this energy contribution is essentially independent of the environment (solvent or protein) and can therefore be added to the corresponding electrostatic energy terms in the protein environment to obtain the correct total free energy.

Vector-Valued Notation of Protonation States. The energy terms introduced above can be used to define the total electrostatic energy for a given protonation pattern of the titratable residues in the protein. However, for this purpose, we first need to define appropriate vector and matrix valued quantities to manage the book keeping of the different μ-states of the titratable residues. We start by introducing for each titratable residue (i) a pair of vectors $\vec{d}(i) = (d_1(i), d_2(i),..., d_{n[d(i)]}(i))$ and $\vec{h}(i) = (h_1(i), h_2(i), ..., h_{n[h(i)]}(i))$ denoting the μ-states occupied in the deprotonated or protonated state, respectively, which are combined into the μ-state vector

$$\vec{p}\,(i) = (\vec{d}(i), \vec{h}(i)) \tag{15}$$

The components of this μ-state vector are zero except for one component with value unity, whose position denotes the precise protonation state and μ-state of residue *i*. These vectors are then combined into a protonation supervector, the global protonation state vector, denoting protonation states and occupation of the μ-states of all N_{titr} titratable residues

$$\vec{P} = (\vec{p}\,(1), \vec{p}\,(2), ..., \vec{p}\,(N_{\mathrm{titr}})) \tag{16}$$

Vector- and Matrix-Valued Notation of Electrostatic Energy Terms. We also need to define appropriate vector- and matrix-valued quantities for the energy terms to control the book keeping of the many different protonation and μ-states. We first define the interaction matrix of residue-pairs in the protein dielectric environment

$$\mathbf{W}_{\mathrm{two}}^{\mathrm{protein}} = \begin{pmatrix} \mathbf{0} & \mathbf{W}_{1,2} & ... & \mathbf{W}_{1,N_{\mathrm{titr}}} \\ \mathbf{W}_{2,1} & \mathbf{0} & ... & \mathbf{W}_{2,N_{\mathrm{titr}}} \\ .. & .. & .. & .. \\ \mathbf{W}_{N_{\mathrm{titr}},1} & \mathbf{W}_{N_{\mathrm{titr}},2} & ... & \mathbf{0} \end{pmatrix} \tag{17}$$

The individual matrices $\mathbf{W}_{i,j}$ in eq 17 describe the electrostatic interactions between the residues i and j in different μ-states. The matrices $\mathbf{W}_{i,i}$ in the diagonal of $\mathbf{W}_{\mathrm{two}}^{\mathrm{protein}}$ are zero matrices. They would correspond to the self-energy of titratable residues, which are considered in the Born terms of one-residue interactions, eq 11. The matrices $\mathbf{W}_{i,j}$ are defined in accordance with the definition of the μ-state vector, eq 15, accounting for the protonation states of residues i and j

$$\mathbf{W}_{i,j} = \begin{pmatrix} \mathbf{w}_{i,j}[\vec{d}(i),\vec{d}(j)] & \mathbf{w}_{i,j}[\vec{d}(i),\vec{h}(j)] \\ \mathbf{w}_{i,j}[\vec{h}(i),\vec{d}(j)] & \mathbf{w}_{i,j}[\vec{h}(i),\vec{h}(j)] \end{pmatrix} \tag{18}$$

The elements of the matrix $\mathbf{W}_{i,j}$, eq 18, are, themselves, matrices describing the multitude of possible electrostatic interactions between the different μ-states of residues i and j for a fixed protonation pattern (α,β)

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

Table 2. Root-Mean-Square Deviations (RMSD) between Measured and Computed pK_a Values (pK_a RMSD) of 194 Titratable Residues in 13 Proteins in the Benchmark Set Given in pH Units[a]

					RMSD between measured and computed pK_a					
no.	protein name	PDB ID	no. of pK_a[b]	no. of atoms[c]	KB$^+$	PropKa[d]	KB2$^+$MD	KB2$^+$MD Rashin	KB2$^+$MD ε_{min} = 1	KB2$^+$MD ε_{min} = 4
1	pancrea trypsin inhibitor	4PTI	14	892	0.95	0.41	0.66	0.40	0.94	0.65
2	streptococcal protein G	1PGA	15	855	0.94	0.68	0.87	0.50	0.91	0.58
3	Barnase	1A2P	12[e]	1727	1.03	1.17	0.92	0.66	0.73	0.68
4	Lysozyme	2LZT	20	1960	1.08	0.66	0.96	1.07	0.94	0.92
5	bovine ribonuclease A	3RN3	14[f]	1856	0.77	0.79	0.79	0.58	0.77	0.61
6	ribonuclease H	2RN2	20[g]	2455	1.67	0.78	1.04	0.60	1.11	0.84
7	rat T-lymphocyte adhesion glycoprotein	1HNG	14	1576	1.09	0.54	0.89	0.66	0.89	0.81
8	Ca binding protein	3ICB	19	1202	0.95	0.66	0.64	0.66	0.59	0.70
9	ovomucoid inhib OMTKY3	1PPF	11	418	0.84	0.54	0.65	0.60	0.63	0.65
10	thioredoxin	1ERT	17	821	1.22	0.93	1.60	1.39	1.84	1.16
11	xylanase[h]	1XNB	7	1448	1.47	0.89	1.08	1.04	0.98	1.13
12	GCN4 leucine zipper[i]	2ZTA	16/16	1120	1.22/1.13	0.46/0.51	1.12/1.19	0.71/0.78	1.07/1.12	0.80/0.85
13	staphylococ. nucl. (Δ+PHS)	3BDC	15	2101	1.65	0.83	0.61	0.93	0.60	0.50
	all titratable residues		**194**		**1.17**	**0.74**	**0.96**	**0.81**	**0.99**	**0.79**

[a]The pK_a values obtained with two types of electrostatic energy computation (KB$^+$ and KB2$^+$MD) and the empirical approach PropKa 3.1[54,55] are compared. In addition, results for two variants applied to KB2$^+$MD are shown: (i) a different set of vdW radii for electrostatic energy computations (Rashin) and (ii) an energy minimization applied to the MD structures prior to electrostatic energy computations with dielectric constant of ε_{min} = 1 and ε_{min} = 4. [b]Number of pK_a values considered per protein [c]Number of atoms including hydrogens [d]The proteins with the running numbers 2−6, 8, 9, and 11 were used to optimize the parameters of PropKa.[54] [e]Asp75 was not included, as in ref 29. The computed pK_a value of this buried residue is very acidic, whereas the measured value of 3.1 is obtained under unfolding conditions and is therefore close to the value in aqueous solution.[70] [f]His48 was not considered, as in ref 29, since together with Gln101 it undergoes a local but significant conformational change when titrated. [g]Asp10 was not considered, as in ref 29, since it participates in a Mg^{2+} binding site. [h]For xylanase, the pK_a values are computed with the O and N atoms of Asn35 interchanged, as discussed in the text. [i]The PDB structure for the leucine zipper is dimeric. The number of atoms refers to the whole dimer. The pK_a RMSDs of the two monomers are separated by a slash (/). For the overall pK_a RMSD, the pK_a values from both chains are averaged.

$$\mathbf{w}_{i,j}[\vec{\alpha}(i),\vec{\beta}(j)] = \begin{pmatrix} W_{i,j}[\alpha_1(i),\beta_1(j)] & W_{i,j}[\alpha_1(i),\beta_2(j)] & .. & W_{i,j}[\alpha_1(i),\beta_{n[\beta(j)]}(j)] \\ W_{i,j}[\alpha_2(i),\beta_1(j)] & W_{i,j}[\alpha_2(i),\beta_2(j)] & .. & W_{i,j}[\alpha_2(i),\beta_{n[\beta(j)]}(j)] \\ .. & .. & .. & .. \\ W_{i,j}[\alpha_{n[\alpha(i)]}(i),\beta_1(j)] & W_{i,j}[\alpha_1(i),\beta_2(j)] & .. & W_{i,j}[\alpha_{n[\alpha(i)]}(i),\beta_{n[\beta(j)]}(j)] \end{pmatrix} \quad (19)$$

whose matrix elements are evaluated according to eq 8.

Next, we define the vector of one-residue energy terms describing the interaction of all titratable residues in the protein environment with themselves and with protein background charges

$$\vec{G}_{one}^{protein} = (\vec{G}_{1,one}^{protein}, \vec{G}_{2,one}^{protein}, ..., \vec{G}_{N_{titr},one}^{protein}) \quad (20)$$

where

$$\vec{G}_{i,one}^{protein}[\vec{p}(i)] = (\vec{G}_{i,one}^{protein}[\vec{d}(i)], \vec{G}_{i,one}^{protein}[\vec{h}(i)]) \quad (21)$$

describes the interaction of residue i for both protonation states and

$$\vec{G}_{i,one}^{protein}[\vec{\alpha}(i)] = (G_{i,one}^{protein}[\alpha_1(i)], G_{i,one}^{protein}[\alpha_2(i)], ..., G_{i,one}^{protein}[\alpha_{n[\alpha(i)]}(i)]) \quad (22)$$

the interaction of residue i for all μ-states of a specific protonation state α in analogy to the definition of the residue-pair interaction matrices. The components of the vector on the rhs of eq 22 are defined by eq 11.

Total Electrostatic Energy Function of a Protein. With the above terms, we can now define an expression for the pH-dependent total electrostatic energy (tEE) function of a protein with N_{titr} residues, which are in a protonation state given by the protonation supervector $\vec{P}$, eq 16

$$G_{total}(\vec{P},\mathrm{pH}) = G_{conf}^{protein} + \vec{G}_{one}^{protein}\vec{P} + \vec{P}\mathbf{W}_{two}^{protein}\vec{P} + \sum_{i=1}^{N_{titr}} \delta[\vec{p}(i)]\, \Delta G_{nonelectrostatic,p-d}^{solvent}(i,\mathrm{pH}) \quad (23)$$

The tEE G_{total} depends also on $MD_m(t_{frame})$, i.e., the type of MD simulation and trajectory frame via the energy terms $G_{conf}^{protein}$, $\Delta\vec{G}_{one}^{protein}$, and $\mathbf{W}_{two}^{protein}$. The sum of the three terms in the first line in eq 23 is the tEE function of the whole protein with the titratable residues in the protonation state given by the supervector $\vec{P}$. This electrostatic energy expression vanishes if all charges are moved at infinite distances. The term in the second line of eq 23 connects the tEE of the protein with solvent pH and the pK_a values of the titratable residues measured in aqueous solution, using $\delta[\vec{p}(i)]$, which is unity if residue i is protonated according to the global protonation state $\vec{P}$ and zero otherwise. This term accounts for the additional nonelectrostatic energy needed to protonate the titratable residues that are protonated according to the global protonation state vector $\vec{P}$. No thermodynamic cycle is needed to evaluate these electrostatic energy terms.

Averaging Electrostatic Energies over MD Simulation Data. In eq 23, we insert for the three energy terms the thermal averages obtained from the individual MD simulations (MD_m), which are $\langle\Delta G_{conf}\rangle_{MDm}$, $\langle\Delta\vec{G}_{one}\rangle_{MDm}$, and $\langle\mathbf{W}_{two}\rangle_{MDm}$. The resulting tEE $\langle G_{total}(\vec{P},\mathrm{pH})\rangle_{MDm}$ is thus a function of the of the trajectory (MD_m), i.e., the protonation states and μ-states $\vec{P}$. The dependence on the trajectory type is analogous to the usage of pH adapted conformations (PAC) introduced in the KB$^+$ procedure. In the subsequent MC procedure, Boltzmann averages over the different trajectory types and protonation

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

patterns and μ-states are performed as done in previous work of KB$^+$.[29]

Variation of the vdW Radii. The Karlsberg$^+$ program uses atomic radii from the CHARMM22[62] force field to define the solvent excluded surfaces (SES) of protein and titratable residues in solution needed for the electrostatic energy computations. An alternative set of atomic radii was used by Nina et al.[68] Atomic radii introduced by Rashin et al.[69] to find protein cavities were used by Nilsson et al.[41] to define the corresponding SES for the computation of pK_a values of a leucine zipper. These atomic radii are generally smaller than those employed in our applications.[29] This is especially the case for the oxygen atom, whose radius is 1.4 Å according to Rashin et al.,[69] instead of values around 1.7 Å used for the different oxygen atom types in the CHARMM22[62] force field. To test the effect of these differences, we recalculated our results for the same MD structures using the Rashin atomic radii for the SES determination.

Karlsberg2$^+$ Software. The small variations indicated below the procedures [KB2$^+$MD, KB2$^+$MD-Rashin, KB2$^+$MD-(ε_{min}=1), KB2$^+$MD(ε_{min}=4)] introduced above are implemented in the new version of the software KB$^+$: Karlsberg2$^+$ (KB2$^+$). The complete protocol, except for the initial modeling decisions described in step 1, is performed automatically by the software KB2$^+$. This ensures that all proteins in the benchmark set have been treated in exactly the same way and that the procedure can be easily applied to new protein structures for future applications. The KB2$^+$ software is currently still under development and will be published and released in the near future.

RESULTS

Table 2 lists the pK_a RMSDs between measured and computed values. The pK_a values of all 194 individual residues are listed in Tables S4−S16 of the Supporting Information. Results obtained with KB$^+$, the new procedure that uses data from MD simulations (KB2$^+$MD), and PropKa 3.1[54,55] are compared. For the KB2$^+$MD procedure, the structures from the MD simulations are used without postprocessing. The results are listed in Table 2 under the entries KB2$^+$MD and KB2$^+$MD-Rashin, where the molecular surfaces are either based on the atomic radii from the CHARMM22 force field[62] or from Rashin et al.,[69] respectively. In the two right columns of Table 2, the protein structures are energy-minimized before the computation of the electrostatic energies, which is performed in a homogeneous dielectric medium with dielectric constants of 1 or 4. The latter procedure yields the best results with a pK_a RMSD over all residues in the benchmark set of 0.79 as compared to 1.17 obtained with KB$^+$.

The standard KB2$^+$MD procedure provides computed pK_a values whose RSMD (0.96) from measured values is significantly smaller than that of those predicted with the KB$^+$ method (1.17). Using the atomic radii from Rashin et al.[69] decreases the pK_a RMSD further to 0.81. Energy minimizing the trajectory frames in the KB2$^+$MD procedure with a dielectric constant of ε_{min} = 1 [denoted KB2$^+$MD(ε_{min}=1)] yielded no decrease in the pK_a RMSD (0.99), whereas with ε_{min} = 4 [denoted KB2$^+$MD(ε_{min}=4)], the decrease (0.79) is significant. Combining the atomic radii of Rashin with energy minimization at ε_{min} = 4 resulted in no further improvement over using either procedure alone (results not shown). The null model assumes that the pK_a values in the protein are equal to the values of the corresponding compound in aqueous solution; this model yields an overall pK_a RMSD of 0.97. Since the majority of titratable residues in proteins, in fact, exhibit small pK_a shifts relative to aqueous solution, this result is expected. Even in our best procedure [KB2$^+$MD(ε_{min} = 4)] the pK_a RMSDs of different proteins can vary by a factor of 2, which is mainly influenced by outliers present in one protein but not in the other.

To investigate how the length of MD simulations influences the result of the pK_a computations, we compared results obtained with the KB2$^+$MD(ε_{min}=4) method using either all 90 trajectory frames or else the first (frames 10−54) or second halves (frames 55−99) of the MD simulations. The resulting RMSDs between measured and computed pK_a values were 0.79, 0.88, and 0.84, respectively. This demonstrates that simulations of even less than 10 ns can be sufficient to obtain a reasonably good result for pK_a computations.

For comparison, the pK_a values of the benchmark set were also calculated with the widely used empirical pK_a prediction software PropKa 3.1,[54,55] yielding an overall pK_a RMSD of 0.74 (Table 2). The results obtained with PropKa are slightly better than with the KB2$^+$MD(ε_{min}=4) procedure. It should be noted that the pK_a values from 8 of the 13 proteins in the current benchmark set are part of the training set as described in the publication of PropKa.[54]

Figure 1 shows a scatter plot of the differences between measured and computed shifts of pK_a values for the 194

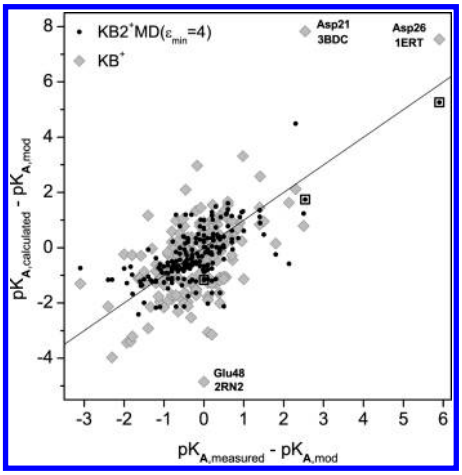


Figure 1. Correlation diagram of measured and computed pK_a shifts relative to the measured pK_a values in solution (p$K_{A,mod}$) for 194 pK_a values of titratable residues in 13 proteins (listed in Table 2). Results obtained with KB$^+$ are shown as gray diamonds. Those computed with KB2$^+$MD(ε_{min}=4) (energy minimization with ε_{min} = 4, see the text) are shown as black filled circles. The diagonal line has a slope of unity and marks perfect agreement between measured and computed results. The most obvious difference between the two computational results is that most of the outliers in the KB$^+$ results are absent in the KB2$^+$MD(ε_{min}=4) results. Three KB$^+$ data points that are strong outliers are labeled. The corresponding shifted values obtained with KB2$^+$MD(ε_{min}=4) are highlighted with open squares.

titratable residues in the 13 different proteins obtained with either the KB$^+$,[29] or KB2$^+$MD(ε_{min}=4) procedure. Using KB2$^+$MD(ε_{min} = 4) instead of KB$^+$ shows a systematic improvement and removal of outliers. The maximum error in any residue is −2.71 (Glu41 in 1HNG) for the new procedure, whereas it was 5.3 (Asp21 in 3BDC) for KB$^+$. The performance of the KB2$^+$MD procedure for the individual residue types is shown in Table 3. The best results were obtained for lysine and

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

Table 3. RMSDs between Computed and Measured pK_a Values for Specific Residue Types[a]

		RMSD between measured and computed pK_a values					
residue type	no. of residues	KB$^+$	PropKa	KB2$^+$MD	KB2$^+$MD Rashin	KB2$^+$MD ϵ_{min} = 1	KB2$^+$MD ϵ_{min} = 4
Asp	55	1.36	0.74	0.98	1.07	1.08	0.75
Glu	73	1.20	0.80	1.13	0.73	1.12	0.91
Lys	32	0.86	0.53	0.59	0.53	0.54	0.54
His	11	0.76	0.98	0.90	0.77	0.87	0.84
Tyr	11	1.03	0.66	0.73	0.70	0.74	0.74
N-ter	4	1.84	0.24	0.79	0.67	1.41	1.15
C-ter	8	0.91	0.65	0.70	0.51	0.70	0.55

[a]N-ter and C-ter are the N- and C-terminus of the protein, respectively. The benchmark set does not contain cysteine and arginine residues, since no measured pK_a values are available for these residues.

the C-termini, with overall pK_a RSMDs of 0.54 and 0.55, respectively, whereas the worst results were for the N-termini, with an overall pK_a RMSD of 1.15. It should be noted here that only 4 pK_a's of N-termini are contained in the benchmark set.

The average coordinate RMSDs of protein backbone relative to the corresponding crystal structures over the MD simulations of the 13 proteins are given in Table S1 of the Supporting Information. For lysozyme, the time evolution of the coordinate RMSDs for all four MD trajectories are plotted in Figure 2. Although the nonequilibrium protonation patterns

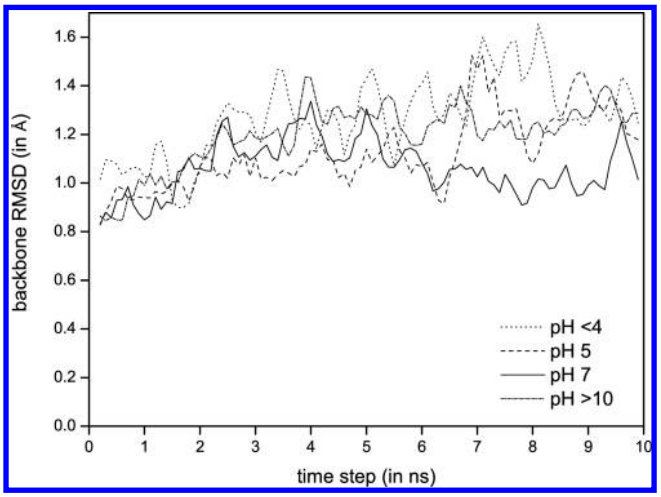


Figure 2. Backbone coordinate RMSDs in the MD simulations for lysozyme relative to the crystal structure (PDB ID 2LZT). In each of the MD simulations, the protein has one of four different protonation patterns, pH < 4, pH 5, pH 7, and pH > 10, defined in Table 1. The curves are running averages over three trajectory frames with 50 ps between frames.

pH < 4 and pH > 10 differ considerably from those for pH 5 and pH 7 (see Table 1), the average coordinate RMSDs of the corresponding trajectories are similar: 1.3 and 1.2 Å for pH < 4 and pH > 10, respectively, compared to 1.1 Å for both pH 5 and pH 7. Hence, coordinate RMSDs are not sensitive to the protonation state of the protein.

DISCUSSION

Total Electrostatic Energy Function for the Computation of pK_a Values in Proteins. The tEE, eq 23, developed in this study differs from the more traditional approaches[22,29] for pK_a computations of titratable residues in proteins where the a so-called intrinsic pK_a value is used.[11] The intrinsic pK_a value describes the pK_a value for each individual titratable residue with all other titratable residues in a fictitious reference protonation state, which usually is the standard protonation state in aqueous solution where bases are protonated and acids are deprotonated. The intrinsic pK_a values are not easily interpretable, since they refer to a specific protonation state that is not necessarily valid and combine information from protein and solvent environments. To describe the energetics of different protonation patterns and to introduce the mutual interaction of titratable residues in different protonation states, a residue-pair interaction is introduced in these methods, which corrects for the arbitrary reference protonation states used for computing the intrinsic pK_a values. A residue-pair interaction vanishes if at least one of the two residues is in the reference protonation state. The total electrostatic energy vanishes if the reference protonation pattern is adopted. Hence, the residue-pair interaction depends strongly on the chosen reference protonation state and is therefore also not easily interpretable. In addition, generalizing this approach to account for different conformations of the protein also requires the definition of an arbitrary reference conformation.

The tEE introduced in the present study employs neither reference protonations states nor reference conformations. Nevertheless, the conformational variability is fully considered using structures from MD simulations with different protonation patterns that are combined by thermodynamic averaging. As is common for electrostatic energy functions, the new function tEE vanishes if all charges are at infinite distances. The first three energy terms in eq 23 are directly interpretable. The term $G_{conf}^{protein}$ describes how the electrostatic energy varies with protein conformation, ignoring the titratable residues. The term $\vec{G}_{one}^{protein}\vec{P}$ describes the electrostatic interaction between titratable residues in protonation state $\vec{P}$, eq 16, and protein conformation. The residue-pair interactions $\vec{P}\mathbf{W}_{two}^{protein}\vec{P}$ describe the mutual interaction of all titratable residues in protonation state $\vec{P}$.

A problem with traditional methods is that to cancel grid artifacts and avoid spurious unphysical contributions from 1 to 2, 1−3, and 1−4 electrostatic atom-pair interactions, the same conformations of the titratable residues are used in the protein and solvent environments. In the current formulation of the electrostatic energy tEE, eq 23, such spurious electrostatic interactions are explicitly removed, as described above in connection with eqs 5, 6, and 9. Hence, the residue conformations in the solvent can differ from those that they adopt in the proteins. It would even allow a separate parametrization of the environment-independent nonelectrostatic energy contribution of the deprotonation reaction, described by the last term in eq 23 but not applied in the present study. However, to facilitate a comparison with the

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

previous methods, in the present study we used the same conformations of titratable residues in both the protein and solvent environments.

Using Different Atomic Radii. Use of the set of atomic radii from Rashin at al.[69] to define the protein volume for electrostatic energy computations instead of the CHARMM22 radii improved the accuracy of the computed pK_a values, lowering the pK_a RMSD from 0.96 to 0.81. This improved result is comparable to the pK_a RMSD of 0.79 obtained with the KB2$^+$MD(ε_{min}=4) procedure. The KB2$^+$MD(ε_{min}=4) procedure yields a maximum error of +2.71 pH units (for Glu41 in 1HNG) as compared to −4.61 pH units (for Asp21 in 3BDC) using the KB2$^+$MD-Rashin procedure. The latter tends to underestimate the pK_a shifts of acids, yielding pK_a values that are closer to the pK_a value in aqueous solution as compared to the KB2$^+$MD(ε_{min}=4) procedure (Table S2 of the Supporting Information). The most significant difference in atomic radii between the CHARMM22 force field and those used by Rashin at al. is for oxygen, whose radius is 0.3 Å smaller in the latter. Hence, the atomic charges of oxygen atoms at the protein surface are closer to the solvent dielectric medium. As a consequence, the pK_a values of the acidic residues (Glu, Asp, and C-ter) are closer to the pK_a value in aqueous solution.

In terms of the average error in pK_a values, there are no significant differences between the two procedures [KB2$^+$MD-(ε_{min}=4) and KB2$^+$MD-Rashin], as shown in Table S3 of the Supporting Information. The CHARMM22[62] radii that are used by KB$^+$ are optimized for MD simulations. The radii taken from Rashin et al.[69] have been optimized to find cavities in protein structures. Neither has been optimized for electrostatic energy computations, but their use and comparison demonstrates the strong influence that differences in atomic radii can have. This suggests a reparameterization of the atomic radii for pK_a calculations. We also combined both procedures [KB2$^+$MD-Rashin(ε_{min}=4)], obtaining an overall RMSD of 0.79 pH units, which is the same value obtained for KB2$^+$MD(ε_{min}=4), demonstrating that the positive effect of the two variations in the procedure is not additive.

Effect of Energy Minimization. Minimizing the energy of the whole protein water box with a dielectric constant of ε_{min} = 4, before electrostatic energy computations were performed, improved the agreement with the measured pK_a values significantly. In contrast, the same energy minimization procedure yielded no improvement if performed with ε_{min} = 1 (see Table 1). An analysis of this dependence can be summarized as follows. Energy minimization removes primarily the influence of kinetic energy on protein structures obtained by MD simulation. Thus, it regularizes structures. However, in the absence of kinetic energy, attractive electrostatic interactions are emphasized too much. As a consequence, H-bonds and salt bridges that stabilize specific protonation patterns are strengthened. For atom pairs, where the electrostatic interactions are attractive, the Lennard-Jones (LJ) potentials that model the vdW interactions are generally repulsive. If the energy minimization is performed with weaker electrostatic interactions (ε_{min} = 4), the relative influence of the LJ interactions is enhanced. As a consequence, H-bonds and salt bridges are weakened, leading to a more balanced regularization of the structures.

Role of MD Simulations with Different Protonation Pattern. The accuracy of computed pK_a values can be enhanced if MD simulations are used to include conformational variability. However, conventional MD simulations need to use a specific fixed choice of protonation pattern for all residues. An MD trajectory relaxes around this protonation pattern, and as a consequence, this pattern appears to be more stable than others not used for the MD simulation. Therefore, MD simulations with different sets of protonation patterns are needed to compensate for this bias. For the same reason, different PAC's are used in KB$^+$.[29] We obtained good results with the set of four protonation patterns listed in Table 1, which are a combination of assigning protonation states according to individual-residue pK_a's and according to the results of pK_a computations based on the crystal structures. For example, the acidic residues Asp and Glu are all protonated in MD simulations with pH < 4 and deprotonated in MD simulations with pH 7 and pH > 10, whereas their protonation states are determined by pK_a computations using the crystal structures in MD simulations with pH 5. The basic residue Lys is protonated in MD simulations with pH < 4 and pH 5 and deprotonated with pH > 10, whereas the protonation state is determined by pK_a computations with pH 7. All other residues considered to be titratable (His, Tyr, Cys, Nter, Cter) are protonated for MD simulations with pH < 4 and deprotonated for those with pH > 10, whereas their protonation is determined by pK_a computations for moderate pH values of pH 5 and pH 7. Hence, the accuracy of the final pK_a results in the current study depends not only on the appropriateness of the protonation pattern used for the MD simulations (Table 1) but also indirectly on the accuracy of the initial pK_a computations using the crystal structures.

Empirical versus Electrostatic Energy Based pK_a Value Computation. PropKa[54,55] is an empirical method for predicting pK_a values in proteins using a physicochemically motivated parametrization. The method is fast, widely used, and has been improved considerably over the years. In contrast to earlier versions,[29] the present version of PropKa, 3.1, also yields good results for protein residue pK_a values that are shifted considerably compared to those in aqueous solution. It was demonstrated that the new PropKa 3.1 yields significantly better agreement with measured pK_a values in proteins than methods based on electrostatic energy computations.[54,55]

Among the 194 titratable residues with measured pK_a values, there are 38 residues whose side chains are buried to more than 80%, where the degree of being buried is evaluated by computing the solvent exclusion surface of the residue side chains. Among these 38 buried residues are 17 Glu, 15 Asp, 5 Tyr, and 1 His. These numbers correspond to the occurrences in the whole data set of 194 measured titratable residues except for tyrosine. There are 11 tyrosines in the whole data set, of which 5 are buried. Since tyrosine is hydrophobic, it has a tendency to be buried. Measured pK_a shifts between protein and solvent environments are generally larger for buried residues than for solvent exposed residues. The 38 buried residues exhibit an average measured pK_a shift of $\langle|\Delta pK_a|\rangle$ = 1.13, as compared to $\langle|\Delta pK_a|\rangle$ = 0.54 for the 156 less or not buried titratable residues. The pK_a RMSD is for the KB2$^+$MD(ε_{min}=4) approach in this work (see Table 2 for a description), 1.08 and 0.70 for the buried and less buried residues, respectively. The corresponding values for PropKa 3.1 are 1.15 and 0.59. These numbers differ more for PropKa than with the present approach. PropKa is an empirical method with physically motivated terms that are parametrized by using known experimental data, which works better if more data are available.

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

In the present approach, by combining electrostatic energy with MD simulation, the quality of agreement with measured values is nearly equal to PropKa (Table 2). However, the present electrostatic approach is quite expensive. Hence, for routine pK_a predictions in uncomplicated or straightforward cases, the use of PropKa[54,55] by itself is generally adequate and often preferred.

However, for non-routine cases, i.e., those that are more complex, unusual, or otherwise do not lend themselves to an empirical approach, a more detailed, physics-based approach like the current one becomes necessary. Non-routine cases involving pK_a computations include titratable residues that are in contact with redox active centers, such as the two propionates bound to heme,[71] or reaction centers like the Mn-cluster in photosystem II (PSII).[72] Other non-routine cases are non-equilibrium scenarios, which occur for proton conduction, creation, and depletion processes. Prominent examples are processes of proton creation and conduction at the Mn-cluster of PSII[73] or proton conduction and depletion in cytochrome c oxidase.[74] In the latter case, it may be necessary to introduce an additional FPP with a protonation pattern that differs from the four listed in Table 1, for which a corresponding MD simulation is necessary to account for unconventional protonation states. Another advantage of the electrostatic energy approach over empirical methods is the option to analyze the problem in terms of energy contributions and individual protein conformations to determine which ones contribute preferentially to the computed pK_a value. Also, a particular advantage of the approach described in the current article is the possibility for *ab initio* computation of pK_a values for arbitrary organic compounds. Here, high-level quantum chemistry calculations would need to be combined with the electrostatic energy computations.[9]

Future Developments. The method presented here is limited to structural changes in proteins that can be generated with classical MD simulations. The results indicate that standard protein MD simulations over a 10 ns time span are sufficient for computing reliable pK_a values. If required by the particular problem, then the protocol can be easily changed using longer or even alternative MD simulations with different protonation states and even different MD simulation techniques. However, a requirement is that the MD simulations for different protonation patterns of the protein generate thermodynamically correct ensembles.

The Karlsberg$^+$ (KB^+) program uses local structural relaxation involving amino acid side chains to adapt the protein conformations to different pH values thus yielding the PACs.[29] Using structures from MD simulations to compute pK_a values in proteins with the Karlsberg$^+$ machinery, as was done in the present study, is a novel approach. The proteins used to test the new method do not form complexes with ligands, ions, or other proteins. Some of the structures have binding pockets for ions (e.g., staphylococcal nuclease binds calcium[57]), but the related pK_a values have been measured in their absence. However, with a straightforward extension of the present procedure, it would also be possible to compute pK_a values in proteins involving bound ligands, ions, or other cofactors, although computing pH-dependent binding affinities may introduce additional complications.[75]

The difficulty of calculating pK_a values with electrostatic energy evaluations in staphylococcal nuclease mutants has been discussed extensively in a recent work.[76] Two main problems have been described: (i) significant structural changes caused by water molecules entering the protein interior in the neighborhood of charged residues and (ii) inclusion of water molecules buried in small hydrophilic cavities. For the latter problem, we proposed a solution that describes buried waters implicitly by locating internal cavities and filling their volume with a dielectric continuum with a large dielectric constant.[76] A solution for the first problem could be the use of sufficiently long MD simulations to sample the conformation changes.[77] Furthermore, combining improved methods for describing small internal cavities with proper sampling of larger structural changes using MD simulation, as described here, could help to address challenging pK_a computations in proteins containing buried charged amino acids not involved in salt bridges, such as also occur in the mutants of staphylococcal nuclease.[78]

CONCLUSIONS

In the present study, structures from MD simulations with different specific protonation patterns are used to compute pK_a values of titratable groups in proteins using electrostatic energies obtained by solving the Poisson−Boltzmann equation. The work is focused on an automatic procedure that should be valid for a large number of pK_a values in different proteins. When considering individual titratable residues that are of special functional interest, more detailed optimization can be performed. With this approach, the agreement between computed and measured pK_a values is shown to be significantly improved compared to that with the KB^{+},[29] method, which essentially uses protein crystal structure information combined only with local structural variation of the side chains of titratable residues. Performing energy minimization of the individual MD structures with $\varepsilon_{min} = 4$ prior to electrostatic energy evaluation yields the best results, with an pK_a RMSD of 0.79 pH units over the entire set of 194 measured pK_a values in 13 proteins. The reliability in terms of the maximum deviation between computed and measured pK_a values is improved, since outliers are avoided with the new method. It is also demonstrated here that with a set of atomic radii, which differs from the vdW radii in the CHARMM22[62] force field conventionally used in our approach, the pK_a RMSD value can be lowered in the $KB2^{+}$MD approach if no energy minimization is used prior to electrostatic energy computation. However, combining the use of the different atomic radii with energy minimization did not yield additional improvement. This indicates that a careful reparameterization of atomic partial charges and radii may improve the agreement between computed and measured pK_a values in proteins, which will be the subject of future work.

The present procedure is considerably more CPU time-intensive, since four MD simulations, with a combined length of 40 ns, and 720 electrostatic energy computations (4×90 TAPBS[29] + 4×90 APBS[66]) are required for each protein. Sampling protein conformations is performed with the well-established MD simulation technique, which is a huge advantage over elaborate modeling methods. To compute thermodynamically averaged pK_a values from the different MD simulations, the Boltzmann factors corresponding to the total electrostatic energies of protein conformations and protonation patterns of the individual MD trajectories are used. These Boltzmann factors can also be used to identify and analyze the protein conformations that prevail at specific pH values. Programs that perform such computations will be made available soon.

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

The current work also introduces a new expression for the electrostatic energy function, in which no thermodynamic cycle and no reference conformation and protonation are needed. As a consequence, (i) the electrostatic energy terms referring to protein and solvent are separated, (ii) the pair interactions of titratable residues are easily interpretable, (iii) the conformations of titratable residues in solvent and protein environments can differ, and (iv) instead of computing the electrostatic solvation energy of the titratable residues in aqueous solution, these energies could be parametrized for each residue type. The latter point may allow for more accurate pK_a calculations in the future.

■ ASSOCIATED CONTENT

Supporting Information

Table S1: Average backbone RMSD values of all MD trajectories. Table S2: Shift between calculated and model pK_a value by residue type. Table S3: Shift between calculated and experimental pK_a values by residue type. Tables S4−S16: Individual experimental and calculated pK_a values for all 194 titratable residues. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00123.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: knapp@chemie.fu-berlin.de.

Funding

The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG) in the frame of the collaborative research centers (CRC) 1078 and 765 with projects C2 and C1, respectively.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Milan Hodoscek for useful discussions and Dr. Robert Petrella for critical reading of the manuscript.

■ ABBREVIATIONS

APBS, adaptive Poisson−Boltzmann solver; FPP, fixed protonation pattern; KB, karlsberg; MD, molecular dynamics; LJ, Lennard-Jones; *l*PBE, linearized Poisson−Boltzmann equation; MC, Monte Carlo; μ-state, protonation microstate; PAC, pH adapted conformer; PDB, Protein Data Bank; pK_a-RMSD, pK_a root-mean-square deviation; SES, solvent excluded surface; tEE, total electrostatic energy; vdW, van der Waals

■ REFERENCES

(1) Warshel, A. Calculations of Enzymatic Reactions: Calculations of pK_a, Proton Transfer Reactions, and General Acid Catalysis Reactions in Enzymes. *Biochemistry* **1981**, *20*, 3167−3177.

(2) Warshel, A.; Åqvist, J. Electrostatic Energy and Macromolecular Function. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 267−298.

(3) McDonald, I. K.; Thornton, J. M. Satisfying Hydrogen Bonding Potential in Proteins. *J. Mol. Biol.* **1994**, *238*, 777−793.

(4) Honig, B.; Nicholls, A. Classical Electrostatics in Biology and Chemistry. *Science* **1995**, *268*, 1144−1149.

(5) Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. Protein Ionizable Groups: pK_a Values and Their Contribution to Protein Stability and Solubility. *J. Biol. Chem.* **2009**, *284*, 13285−13289.

(6) Yam, R.; Nachliel, E.; Gutman, M. Time-Resolved Proton Protein Interaction. Methodology and Kinetic Analysis. *J. Am. Chem. Soc.* **1988**, *110*, 2636−2640.

(7) Zscherp, C.; Schlesinger, R.; Tittor, J.; Oesterhelt, D.; Heberle, J. In Situ Determination of Transient pK_a Changes of Internal Amino Acids of Bacteriorhodopsin by Using Time-Resolved Attenuated Total Reflection Fourier-Transform Infrared Spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5498−5503.

(8) Nielsen, J. E.; Gunner, M. R.; García-Moreno, E. B. The pK_a Cooperative: A Collaborative Effort to Advance Structure-Based Calculations of pK_a Values and Electrostatic Effects in Proteins. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3249−3259.

(9) Schmidt am Busch, M.; Knapp, E. W. Accurate pK_a Determination for a Heterogeneous Group of Organic Molecules. *ChemPhysChem* **2004**, *5*, 1513−1522.

(10) Galstyan, G.; Knapp, E. W. Computing pK_a Values of Hexa-aqua Transition Metal Complexes. *J. Comput. Chem.* **2015**, *36*, 69−78.

(11) Tanford, C.; Kirkwood, J. G. Theory of Protein Titration Curves. *J. Am. Chem. Soc.* **1957**, *79*, 5333−5339.

(12) Warshel, A.; Russell, S. T. Calculations of Electrostatic Interactions in Biological Systems and in Solutions. *Q. Rev. Biophys.* **1984**, *17*, 283−422.

(13) Bashford, D.; Karplus, M. pK_a's of Ionizable Groups in Proteins: Atomic Detail from a Continuum Electrostatic Model. *Biochemistry* **1990**, *29*, 10219−10225.

(14) Bashford, D. Electrostatic Effects in Biological Molecules. *Curr. Opin. Struct. Biol.* **1991**, *1*, 175−184.

(15) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. Protonation of Interacting Residues in a Protein by a Monte Carlo Method—Application to Lysozyme and the Photosynthetic Reaction Center of *Rhodobacter sphaeroides*. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5804−5808.

(16) Gunner, M. R.; Honig, B. Electrostatic Control of Midpoint Potentials in the Cytochrome Subunit of the *Rhodopseudomonas viridis* Reaction Center. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 9151−9155.

(17) Karshikoff, A. A Simple Algorithm for the Calculation of Multiple Site Titration Curves. *Protein Eng.* **1995**, *8*, 243−248.

(18) Demchuk, E.; Wade, R. C. Improving the Continuum Dielectric Approach to Calculating pK_as of Ionizable Groups in Proteins. *J. Phys. Chem.* **1996**, *100*, 17373−17387.

(19) Mehler, E. L. Self-Consistent, Free Energy Based Approximation To Calculate pH Dependent Electrostatic Effects in Proteins. *J. Phys. Chem.* **1996**, *100*, 16006−16018.

(20) Sham, Y. Y.; Chu, Z. T.; Warshel, A. Consistent Calculations of pK_a's of Ionizable Residues in Proteins: Semi-microscopic and Microscopic Approaches. *J. Phys. Chem. B* **1997**, *101*, 4458−4472.

(21) Schaefer, M.; Sommer, M.; Karplus, M. pH-Dependence of Protein Stability: Absolute Electrostatic Free Energy Differences between Conformations. *J. Phys. Chem. B* **1997**, *101*, 1663−1683.

(22) Ullmann, G. M.; Knapp, E. W. Electrostatic Models for Computing Protonation and Redox Equilibria in Proteins. *Eur. Biophys. J.* **1999**, *28*, 533−551.

(23) Simonson, T.; Carlsson, J.; Case, D. A. Proton Binding to Proteins: pK_a Calculations with Explicit and Implicit Solvent Models. *J. Am. Chem. Soc.* **2004**, *126*, 4167−4180.

(24) Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the Prediction of pK_a Values in Proteins. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3260−3275.

(25) Ullmann, G. M.; Bombarda, E. pK_a Values and Redox Potentials of Proteins. What Do They Mean? *Biol. Chem.* **2013**, *394*, 611.

(26) Lin, Y.-L.; Aleksandrov, A.; Simonson, T.; Roux, B. An Overview of Electrostatic Free Energy Computations for Solutions and Proteins. *J. Chem. Theory Comput.* **2014**, *10*, 2690−2709.

(27) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

(28) Rabenstein, B.; Knapp, E. W. Calculated pH-Dependent Population and Protonation of Carbon-Monoxy-Myoglobin Conformers. *Biophys. J.* **2001**, *80*, 1141−1150.
(29) Kieseritzky, G.; Knapp, E. W. Optimizing pK_A Computation in Proteins with pH Adapted Conformations. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1335−1348.
(30) Sandberg, L.; Edholm, O. pK_a Calculations Along a Bacteriorhodopsin Molecular Dynamics Trajectory. *Biophys. Chem.* **1997**, *65*, 189−204.
(31) Zhou, H.-X.; Vijayakumar, M. Modeling of Protein Conformational Fluctuations in pK_a Predictions. *J. Mol. Biol.* **1997**, *267*, 1002−1011.
(32) Wlodek, S. T.; Antosiewicz, J.; McCammon, J. A. Prediction of Titration Properties of Structures of a Protein Derived from Molecular Dynamics Trajectories. *Protein Sci.* **1997**, *6*, 373−382.
(33) van Vlijmen, H. W. T.; Schaefer, M.; Karplus, M. Improving the Accuracy of Protein pK_a Calculations: Conformational Averaging versus the Average Structure. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 145−158.
(34) Koumanov, A.; Karshikoff, A.; Friis, E. P.; Borchert, T. V. Conformational Averaging in p*K* Calculations: Improvement and Limitations in Prediction of Ionization Properties of Proteins. *J. Phys. Chem. B* **2001**, *105*, 9339−9344.
(35) Alexov, E. Role of the Protein Side-Chain Fluctuations on the Strength of Pair-Wise Electrostatic Interactions: Comparing Experimental with Computed pK_as. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 94−103.
(36) Nielsen, J. E.; McCammon, J. A. On the Evaluation and Optimization of Protein X-ray Structures for pK_a Calculations. *Protein Sci.* **2003**, *12*, 313−326.
(37) Eberini, I.; Baptista, A. M.; Gianazza, E.; Fraternali, F.; Beringhelli, T. Reorganization in Apo- and Holo-β-Lactoglobulin Upon Protonation of Glu89: Molecular Dynamics and pK_a Calculations. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 744−758.
(38) Kuhn, B.; Kollman, P. A.; Stahl, M. Prediction of pK_a Shifts in Proteins Using a Combination of Molecular Mechanical and Continuum Solvent Calculations. *J. Comput. Chem.* **2004**, *25*, 1865−1872.
(39) Archontis, G.; Simonson, T. Proton Binding to Proteins: A Free-Energy Component Analysis Using a Dielectric Continuum Model. *Biophys. J.* **2005**, *88*, 3888−3904.
(40) Makowska, J.; Baginska, K.; Makowski, M.; Jagielska, A.; Liwo, A.; Kasprzykowski, F.; Chmurzynski, L.; Scheraga, H. A. Assessment of Two Theoretical Methods to Estimate Potentiometric Titration Curves of Peptides: Comparison with Experiment. *J. Phys. Chem. B* **2006**, *110*, 4451−4458.
(41) Nilsson, L.; Karshikoff, A. Multiple pH Regime Molecular Dynamics Simulation for p*K* Calculations. *PLoS One* **2011**, *6*, e20116.
(42) Mertz, J. E.; Pettitt, B. M. Molecular-Dynamics at a Constant pH. *Int. J. Supercomput. Appl. High Perform. Comput.* **1994**, *8*, 47−53.
(43) Baptista, A. M.; Martel, P. J.; Petersen, S. B. Simulation of Protein Conformational Freedom as a Function of pH: Constant-pH Molecular Dynamics Using Implicit Titration. *Proteins: Struct., Funct., Genet.* **1997**, *27*, 523−544.
(44) Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. Simulating Proteins at Constant pH: An Approach Combining Molecular Dynamics and Monte Carlo Simulation. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 469−480.
(45) Lee, M. S.; Freddie R Salsbury, J.; Brooks, C. L., III Constant-pH Molecular Dynamics Using Continuous Titration Coordinates. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738−752.
(46) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* **2004**, *25*, 2038−2048.
(47) Machuqueiro, M.; Baptista, A. M. Constant-pH Molecular Dynamics with Ionic Strength Effects: Protonation Conformation Coupling in Decalysine. *J. Phys. Chem. B* **2006**, *110*, 2927−2933.
(48) Stern, H. A. Molecular Simulation with Variable Protonation States at Constant pH. *J. Chem. Phys.* **2007**, 126.
(49) Machuqueiro, M.; Baptista, A. M. Acidic Range Titration of HEWL Using a Constant-pH Molecular Dynamics Method. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 289−298.
(50) Wallace, J. A.; Shen, J. K. Predicting pK_a Values with Continuous Constant pH Molecular Dynamics. *Methods Enzymol.* **2009**, *466*, 455−475.
(51) Foit, L.; George, J. S.; Zhang, B. W.; Brooks, C. L., III; Bardwell, J. C. A. Chaperone Activation by Unfolding. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, E1254−E1262.
(52) Zeng, X.; Mukhopadhyay, S.; Brooks, C. L., III Residue-Level Resolution of Alphavirus Envelope Protein Interactions in pH-Dependent Fusion. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 2034−2039.
(53) Chen, J. H.; Brooks, C. L., III; Khandogin, J. Recent Advances in Implicit Solvent-Based Methods for Biomolecular Simulations. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140−148.
(54) Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.
(55) Sondergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pK_a Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284−2295.
(56) Oshea, E. K.; Klemm, J. D.; Kim, P. S.; Alber, T. X-ray Structure of the GCN4 Leucine Zipper, a 2-Stranded, Parallel Coiled Coil. *Science* **1991**, *254*, 539−544.
(57) Castaneda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; Garcia-Moreno, B. E. Molecular Determinants of the pK_a Values of Asp and Glu Residues in Staphylococcal Nuclease. *Proteins: Struct., Funct., Bioinf.* **2009**, *77*, 570−588.
(58) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining Conformational Flexibility and Continuum Electrostatics for Calculation pK_as in Proteins. *Biophys. J.* **2002**, *83*, 1731−1748.
(59) Joshi, M. D.; Hedberg, A.; McIntosh, L. P. Complete Measurement of the pK_a Values of the Carboxyl and Imidazole Groups in *Bacillus circulans* Xylanase. *Protein Sci.* **1997**, *6*, 2667−2670.
(60) Matousek, W. M.; Ciani, B.; Fitch, C. A.; Garcia-Moreno, E. B.; Kammerer, R. A.; Alexandrescu, A. T. Electrostatic Contributions to the Stability of the GCN4 Leucine Zipper Structure. *J. Mol. Biol.* **2007**, *374*, 206−219.
(61) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy Minimization and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(62) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Hydrogen Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins using the CHARMM22 Force Field. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(63) Im, W.; Lee, M. S.; Brooks, C. L., III Generalized Born Model with a Simple Smoothing Function. *J. Comput. Chem.* **2003**, *24*, 1691−1702.
(64) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(65) Nose, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984**, *81*, 511−519.
(66) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.
(67) Sakalli, I.; Schöberl, J.; Knapp, E. W. mFES: A Robust Molecular Finite Element Solver for Electrostatic Energy Computations. *J. Chem. Theory Comput.* **2014**, *10*, 5095−5112.

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840

(68) Nina, M.; Beglov, D.; Roux, B. Atomic Radii for Continuum Electrostatics Calculations Based on Molecular Dynamics Free Energy Simulations. *J. Phys. Chem. B* **1997**, *101*, 5239−5248.
(69) Rashin, A. A.; Iofin, M.; Honig, B. Internal Cavities and Buried Waters in Globular Proteins. *Biochemistry* **1986**, *25*, 3619−3625.
(70) Oliveberg, M.; Arcus, V. L.; Fersht, A. R. pK_a Values of Carboxyl Groups in the Native and Denatured States of Barnase: The pK_a Values of the Denatured State Are on Average 0.4 Units Lower than Those of Model Compounds. *Biochemistry* **1995**, *34*, 9424−9433.
(71) Gamiz-Hernandez, A. P.; Kieseritzky, G.; Galstyan, A. S.; Demir-Kavuk, O.; Knapp, E. W. Understanding Properties of Cofactors in Proteins: Redox Potentials of Synthetic Cytochromes b. *ChemPhysChem* **2010**, *11*, 1196−1206.
(72) Robertazzi, A.; Galstyan, A.; Knapp, E. W. PSII Manganese Cluster: Protonation of W2, 05, 04 and His337 in the Si State Explored by Combined Quantum Chemical and Electrostatic Energy Computations. *Biochim. Biophys. Acta, Bioenerg.* **2014**, *1837*, 1316−1321.
(73) Ishikita, H.; Saenger, W.; Loll, B.; Biesiadka, J.; Knapp, E. W. Energetics of a Possible Proton Exit Pathway for Water Oxidation in Photosystem II. *Biochemistry* **2006**, *45*, 2063−2071.
(74) Woelke, A. L.; Galstyan, G.; Galstyan, A.; Meyer, T.; Heberle, J.; Knapp, E. W. Exploring the Possible Role of Glu286 in CcO by Electrostatic Energy Computations Combined with Molecular Dynamics. *J. Phys. Chem. B* **2013**, *117*, 12432−12441.
(75) Woelke, A. L.; Kuehne, C.; Meyer, T.; Galstyan, G.; Dernedde, J.; Knapp, E. W. Understanding Selectin Counter-Receptor Binding from Electrostatic Energy Computations and Experimental Binding Studies. *J. Phys. Chem. B* **2013**, *117*, 16443−16454.
(76) Meyer, T.; Kieseritzky, G.; Knapp, E. W. Electrostatic pK_a Computations in Proteins: Role of Internal Cavities. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3320−3332.
(77) Damjanovic, A.; Schlessman, J. L.; Fitch, C. A.; Garcia, A. E.; Garcia-Moreno, B. Role of Flexibility and Polarity as Determinants of the Hydration of Internal Cavities and Pockets in Proteins. *Biophys. J.* **2007**, *93*, 2791−2804.
(78) Harms, M. J.; Castaneda, C. A.; Schlessman, J. L.; Sue, G. R.; Isom, D. G.; Cannon, B. R.; Garcia-Moreno, B. The pK_a Values of Acidic and Basic Residues Buried at the Same Internal Location in a Protein Are Governed by Different Factors. *J. Mol. Biol.* **2009**, *389*, 34−47.

DOI: 10.1021/acs.jctc.5b00123
J. Chem. Theory Comput. 2015, 11, 2827−2840