JCTC Journal of Chemical Theory and Computation

Article

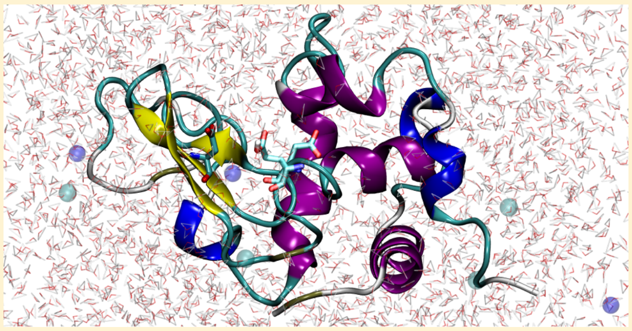
pubs.acs.org/JCTC

Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation

Jason M. Swails,[†] Darrin M. York,[‡] and Adrian E. Roitberg*,[†]

[†]Quantum Theory Project, Chemistry Department, University of Florida, Gainesville, Florida 32611, United States

[‡]BioMaPS Institute for Quantitative Biology, Center for Integrative Proteomics Research, and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08901, United States



ABSTRACT: By utilizing Graphics Processing Units, we show that constant pH molecular dynamics simulations (CpHMD) run in Generalized Born (GB) implicit solvent for long time scales can yield poor pK_a predictions as a result of sampling unrealistic conformations. To address this shortcoming, we present a method for performing constant pH molecular dynamics simulations (CpHMD) in explicit solvent using a discrete protonation state model. The method involves standard molecular dynamics (MD) being propagated in explicit solvent followed by protonation state changes being attempted in GB implicit solvent at fixed intervals. Replica exchange along the pH-dimension (pH-REMD) helps to obtain acceptable titration behavior with the proposed method. We analyzed the effects of various parameters and settings on the titration behavior of CpHMD and pH-REMD in explicit solvent, including the size of the simulation unit cell and the length of the relaxation dynamics following protonation state changes. We tested the method with the amino acid model compounds, a small pentapeptide with two titratable sites, and hen egg white lysozyme (HEWL). The proposed method yields superior predicted pK_a values for HEWL over hundreds of nanoseconds of simulation relative to corresponding predicted values from simulations run in implicit solvent.

1. INTRODUCTION

Solution pH often has a dramatic impact on biomolecular systems. By modulating the protonation state equilibria of various, *titratable* functional groups present in the system, small changes in solution pH can affect the charge distribution within the biomolecule. This charge distribution, in turn, often has a profound impact on the fundamental structure and function of biomolecules. This effect can be so pronounced that some proteins' native states are stable only in a narrow pH range, even denaturing completely in extreme pH environments.[1,2]

Because biomolecular behavior can depend very strongly on the pH-dependent protonation states of various titratable residues, accurate computational models designed to treat such systems must somehow account for pH effects. While the traditional approach of assigning a fixed protonation state for each titratable residue at the beginning of the simulation is still the most common approach, numerous methods have been developed in an attempt to treat pH effects in biomolecules more quantitatively.[3]

Of particular interest in this study is the constant pH molecular dynamics (CpHMD) technique, of which there are several variants.[4−8] CpHMD is a method that leverages the ability of classical molecular dynamics to sample conformational space while simultaneously sampling from the available protonation states according to the semigrand canonical ensemble.[5] By adopting this approach, CpHMD simulations can overcome the limitations imposed by constant-protonation MD simulations by constructing an ensemble whose protonation state distributions are properly weighted via the thermodynamic constraint imposed by a constant chemical potential of hydronium ions.

Within the myriad of available CpHMD methods, there are two fundamentally different approaches—continuous protonation states[6,8−11] and discrete protonation states.[5,7,12−14] In the former, a continuous 'titration coordinate' describing a fictitious 'titration particle' is introduced at each protonable site that is propagated as part of the standard MD according to a pH-dependent force acting on this particle.[6,8] Discrete protonation state models, on the other hand, employ occasional Metropolis Monte Carlo (MC) exchange attempts between different protonation states throughout the course of the MD simulation.[5,7] For the purposes of this study, we will focus on CpHMD models using discrete protonation states—specifically as implemented in the AMBER software suite.[7]

In the original AMBER implementation of CpHMD, the molecular dynamics is propagated treating solvent effects implicitly via the Generalized Born (GB) method.[7] Periodically

Received: December 2, 2013
Published: February 5, 2014

throughout the dynamics, a trial move changing the protonation state of either one or two closely interacting titratable residues is evaluated based on the difference in electrostatic and solvation energies calculated via GB, after which the coordinates are propagated according to this same potential.

Recently, however, Machuqueiro and Baptista raised concerns about pK_a predictions inheriting problems related to the model compound definition and inaccuracies in the underlying force field.[15] In particular, force field deficiencies have been shown to result in incorrect—even unphysical—global minima.[16−20] So far, reported applications of Mongan et al.'s method have shown good results because the simulations were too short to reveal the full extent of the weaknesses in the Generalized Born model being used.[7,12,21−24] When we implemented Mongan et al.'s method on graphics processing units (GPUs) in order to run long simulations, the unphysical states sampled over long time scales degraded the pK_a predictions of the hen egg-white lysozyme (HEWL). This erroneous sampling may be addressed to some extent by using an explicit representation of the solvent to propagate the dynamics.

While most of the physics-based methods designed to describe a biomolecular system at constant pH use an implicit solvent representation of the solvent, several CpHMD methods have been extended to sample, at least conformations, in explicit solvent with both the discrete[5] and continuous[9−11] protonation models. The methods proposed by Baptista et al.[5] and Wallace and Shen[9] use an implicit solvent potential to sample protonation states while the methods developed by Goh et al.[10] and Donnini et al.[11] perform λ-dynamics on the titration coordinate directly in explicit solvent. A more recent approach by Wallace and Shen uses a λ-dynamics approach in pure explicit solvent and adds a counterion whose charge is changed simultaneously with a titratable residue in order to maintain charge neutrality in the unit cell.[25]

Discrete protonation methods use molecular dynamics to propagate the spatial coordinates, while occasionally interrupting the dynamics to attempt changes to the protonation states of the titratable residues using a Metropolis Monte Carlo criteria. The CpHMD method implemented in AMBER[7] (and later implemented in CHARMM[12]) performs MD in GB solvent, periodically attempting to change the protonation state of one or two interacting residues roughly every 10 fs.[7] In the stochastic titration method described by Baptista et al., dynamics is run in explicit solvent for 2 ps,[26] after which a cycle of protonation state change attempts are evaluated using the Poisson−Boltzmann (PB) equation to treat solvation effects for every titratable residue and interacting titratable residue pair. About 40 000 full cycles are attempted each time protonation state changes are attempted.[27] Afterward, the solute is held fixed while MD is propagated on the solvent to reorganize the solvent distribution to the new set of protonation states.

Implicit solvent models—in this case GB and PB—average over all solvent degrees of freedom, thereby instantly incorporating the effects of solvent relaxation around discrete protonation state changes. Therefore, MC moves in which a protonation state change is attempted have a reasonable probability of succeeding when the solution pH is set close to the pK_a of the titratable group. When explicit solvent molecules are present, however, the solvent orientation around any solvent-exposed, titratable residue will oppose any protonation state changes. On average, the solvent distribution tends to resist protonation state changes by imposing a barrier on the order of 100 kcal/mol as estimated by measurements in our lab and in others',[9] making titration with discrete protonation states difficult directly in explicit solvent.

In this study, we present a new method of performing CpHMD simulations in explicit solvent using discrete protonation states. This method is similar in some regards to that used by Baptista et al.,[5] and we evaluate its performance on the model compounds, a pentapeptide, and the HEWL protein. To enhance the sampling capabilities of this new CpHMD method, we use replica exchange in the pH-dimension (pH-REMD), whose theory and performance were discussed previously in the context of implicit solvent calculations.[12,24] This paper is organized as follows: We will first describe the method and its implementation in the Theory and Methods section, followed by a description of the calculations we performed in the Calculation Details section. Afterward, we will evaluate its performance as well as sensitivity to the method's tunable parameters in the Results and Discussion section.

2. THEORY AND METHODS

In this section, we will discuss the details of our proposed method and highlight how it differs from the approach used by Baptista et al.[5] The theoretical foundation of our CpHMD method is described in detail, as well as the pH-REMD method we used in our simulations.

2.1. Conformational and Protonation State Sampling.

In CpHMD, structures are sampled from the semigrand canonical ensemble, whose probability distribution function is given by

$$\rho(\mathbf{q},\, \mathbf{p},\, \mathbf{n}) = \frac{\exp(\beta\mu^{*}n - \beta H(\mathbf{q},\, \mathbf{p},\, \mathbf{n}))}{\sum_{\mathbf{n}'} \int d\mathbf{p}' d\mathbf{q}' \exp(\beta\mu^{*}n' - \beta H(\mathbf{p}',\, \mathbf{q}',\, \mathbf{n}'))} \tag{1}$$

where $\beta = 1/k_BT$, μ^* is the chemical potential of hydronium (directly related to the solution pH), $\mathbf{q}$ is the generalized coordinates of the system particles, $\mathbf{p}$ is the conjugate momenta, and n is the total number of titratable protons present in that state. When bold, $\mathbf{n}$ refers to the protonation state vector, specifying not only the total number of protons present but on which titratable sites those protons are located. The denominator in eq 1 is the *partition function* of the semigrand canonical ensemble.

To sample from the probability function ρ in eq 1, discrete protonation state methods use MD with a fixed set of protonation states to sample coordinates and momenta coupled with a MC-based protonation state sampling at fixed conformations throughout the trajectory. Baptista et al. showed that standard MD, which samples $\rho(\mathbf{q},\mathbf{p}|\mathbf{n})$, used in conjunction with MC moves on protonation states, which samples from $\rho(\mathbf{n}|\mathbf{q},\mathbf{p})$, properly samples from the desired probability distribution function $\rho(\mathbf{q},\, \mathbf{p},\, \mathbf{n})$.[5] In this notation, $\rho(\mathbf{q},\mathbf{p}|\mathbf{n})$ is the conditional probability function of the positions and momenta with fixed protonation states whereas $\rho(\mathbf{n}|\mathbf{q},\mathbf{p})$ is the conditional probability function of the protonation state vector at a fixed protein conformation.

In explicit solvent, $\rho(\mathbf{n}|\mathbf{q},\mathbf{p})$ is difficult to sample directly, since the solvent orientation is relaxed with respect to the current protonation state vector. Following the arguments of Baptista et al., the system coordinates (and momenta) can be separated into solute and solvent degrees of freedom.[5] The protonation state sampling is then performed according to the conditional probability

$$\rho' = \rho(\mathbf{p}_{\text{solvent}}, \mathbf{q}_{\text{solvent}}, \mathbf{n} | \mathbf{p}_{\text{solute}}, \mathbf{q}_{\text{solute}}) \tag{2}$$

where q_{solvent} and p_{solvent} are relaxed solvent distributions of positions and momenta around the protonation state vector, **n**.[5] The energy differences resulting from the relaxed solvent distributions around the different protonation states are quantities that implicit solvent models strive to reproduce. Therefore, the distribution function ρ' in eq 2 can be approximated using continuum models, such as the PB or GB equations, thereby avoiding the otherwise costly solvent relaxation calculation associated with each attempted protonation state change. Performing relaxation MD on the solvent degrees of freedom should be done after protonation state changes to generate the uncorrelated, relaxed solvent distribution required by eq 2. While using MD to generate relaxed solvent distributions violates detailed balance—microscopic reversibility is no longer preserved—Manousiouthakis and Deem have shown that simply satisfying the weaker balance condition is valid.[28]

Contrary to the stochastic titration method that calculated solvation free energies using the PB equation to evaluate protonation state changes,[5] we chose to use the GB implicit solvent model for three main reasons. First, AMBER has numerous GB models readily available,[29−33] allowing us to use the existing code to evaluate protonation state change attempts. Second, results from the original GB-based CpHMD implementation by Mongan et al., and from a number of previous studies using the method, have been promising.[7,22,24,34] Furthermore, GB was shown to be effective when used in a hybrid solvent method with continuous protonation states,[9] is computationally cheaper than PB, and GB is more easily parallelizable, allowing longer simulations to be performed in the same amount of time.

2.2. Explicit Solvent CpHMD Workflow. The process of the CpHMD method presented here can be divided into three repeating steps, summarized in the workflow diagram in Figure 1. This workflow is very similar to the one presented in ref 5 (Figure 2), although the nature of the MC protonation state move is different.

In the proposed method standard MD in explicit solvent is carried out using a constant set of protonation states (an initial set must be provided at the start of the simulation). At some point the MD is stopped, the solvent (including any nonstructural ions) are stripped, the potential is switched to an available GB model, and a set of N protonation state changes are attempted where N is the number of titratable residues. While in principle the MD can be stopped randomly with a predetermined probability at any step, in this iteration of our proposed method we run MD for a set time interval, τ_{MD}, similar to the stochastic titration method.[5]

After the MD is halted and the solvent stripped, protonation state changes are proposed for each titratable residue once, in random order, choosing from the available protonation states of that residue excluding the currently occupied state. The electrostatic energy difference between the proposed and current protonation states, as well as the MC decision regarding whether or not to accept the proposed state, are calculated the same way as in the original GB implementation.[7] If the protonation state change is accepted, the 'current' state is appropriately updated, and the next residue, chosen at random without replacement, is titrated with this new state.

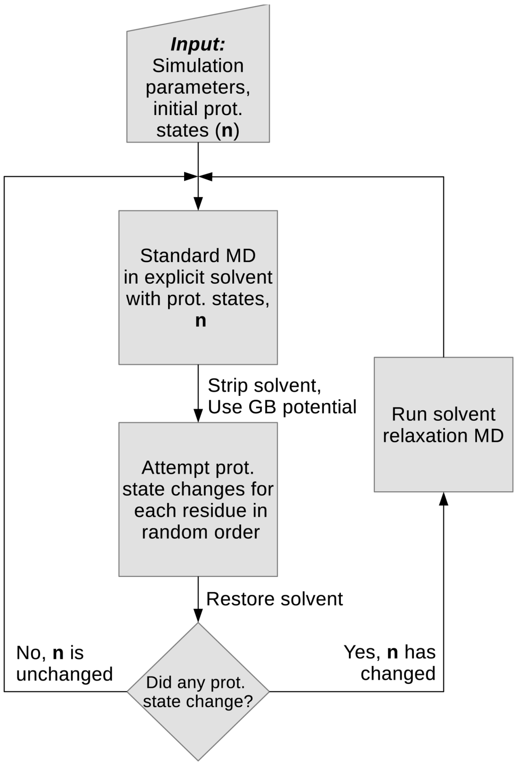

Input:
Simulation parameters, initial prot. states (**n**)
Standard MD in explicit solvent with prot. states, **n**
Strip solvent, Use GB potential
Attempt prot. state changes for each residue in random order
Restore solvent
Did any prot. state change?
No, **n** is unchanged
Yes, **n** has changed
Run solvent relaxation MD

Figure 1. Workflow of the proposed discrete protonation CpHMD method in explicit solvent. Following the standard MD, the solvent, including all nonstructural ions (as determined by user-input), are stripped and the protonation state changes are evaluated in a GB potential. After that, the solvent and the original settings are restored for the remaining steps.

For each residue that is titrated, there is a 25% chance that a so-called multisite titration will occur with a neighboring residue; that is, the proposed change will involve changes to the protonation state of both the residue and its neighbor. Two titratable residues are considered 'neighbors' if any two titrating hydrogen atoms are within 2 Å from each other. If either residue has more than one titrating proton, the two residues are neighbors if the minimum distance between any pair of titrating hydrogens meets the cutoff.

Including multisite protonation state jumps is important for systems that have closely interacting titratable residues. Without these multisite moves, proton transfers between adjacent titratable residues involved in a hydrogen bond would never occur due to the high penalty of disrupting the interaction by adding another proton or removing the proton involved in the hydrogen bond. This feature was actually present in the initial GB CpHMD implementation, and while no mention of it was made in the original paper, a small note was made in the AMBER Users' manual.[7]

If any of the protonation state change attempts were accepted, the solute is frozen while MD is performed on the solvent (and any ions) to relax the solvent distribution around the new protonation states. The length of this relaxation is a tunable parameter of the method, which we will call τ_{rlx}. After the relaxation is complete, the velocities of the solute atoms are restored to their values prior to the relaxation and the standard dynamics is continued.

It is worth noting that as the protonation states change during the course of the CpHMD simulations, so too does the net charge on the system. Because we are using periodic boundary conditions with a lattice-sum method to compute electrostatics, finite-size effects involving the changing charges are introduced.[35,36] The investigations focusing on charge-dependent finite-size effects have identified artifacts affecting primarily computed free energies and pressure and are larger for smaller unit cells.[35] In the proposed method, the protonation state changes are attempted in implicit solvent with a GB potential, which is entirely unaffected by these finite-size effects. Furthermore, since replica exchange simulations in AMBER require the use of constant volume, no pressure corrections are required, either.

2.3. pH-based Replica Exchange. The underlying theory behind replica exchange in pH-space with MD run in explicit solvent is unchanged from the version we implemented in implicit solvent earlier.[12,24] Replicas are ordered by their solution pH parameter, and adjacent replicas attempt to exchange their pH periodically throughout the MD simulations.

The probability of accepting these replica exchange attempts, given by eq 3, depends only on the difference in the number of titrating protons present in each replica and their respective difference in pH.[24]

$$P_{i\to j} = \min\{1, \exp[\ln 10(N_i - N_j)(\mathrm{pH}_i - \mathrm{pH}_j)]\} \tag{3}$$

where N_i is the total number of titratable protons ‘present’ in state *i*.

Because the probability of accepting replica exchange attempts depends only on the number of titratable protons that are present in the system, the number of replicas necessary to obtain efficient mixing in pH-space does not increase as explicit solvent is added. This, coupled with the more accurate sampling found with explicit solvent simulations,[24] makes pH-REMD an effective tool for explicit solvent CpHMD.

3. CALCULATION DETAILS

To evaluate the performance of the proposed method, we applied it to the amino acid model compounds, a small pentapeptide (ACFCA), and a protein commonly used in pK_a calculation studies-the hen egg-white lysozyme (HEWL).

3.1. Implicit Solvent Simulations. In order to allow simulations to be run for hundreds of nanoseconds, we implemented Mongan et al.’s CpHMD method with pH-REMD[24] on GPUs in GB implicit solvent.[37] We used the HEWL structure solved in PDB code 1AKI[38] as our starting structure. The carboxylate residues were initially set in the deprotonated state, and histidine 15 was started in the double-protonated state according to AMBER defaults.

The structure was minimized using 10 steps of the steepest descent algorithm followed by 990 steps of conjugate gradient with 10 kcal/mol·Å^2 . The minimized structure was then heated for 1 ns, varying the target temperature linearly from 10 to 300 K over 667 ps. Positional restraints (5 kcal/mol·Å^2) were placed on the backbone atoms. The temperature was controlled using Langevin dynamics with a 5 ps^{-1} friction coefficient.

The heated structure was then equilibrated at 300 K for 2 ns using Langevin dynamics with a friction coefficient of 10 ps^{-1} with 0.1 kcal/mol·Å^{-2} positional restraints on the backbone. We then ran 500 ns of pH-REMD, attempting to change protonation states every 10 fs and attempting replica exchanges every 20 fs. We used 14 replicas spanning the pH range 1−7.5 with a 0.5 pH-unit spacing between adjacent replicas. No nonbonded cutoff was used for any implicit solvent simulation.

3.2. Model Compounds. Absolute pK_a values are very difficult to calculate in solution; they are impossible using classical force fields. As a result, every physics-based CpHMD method uses the idea of a *model compound* whose experimental pK_a is easy to measure with a high level of accuracy. An empirical parameter—the *reference energy*—is then added so that CpHMD reproduces the experimental pK_a values of these compounds. The reference energy must be set for the solvation model that is used during the simulations, which was the same GB^{OBC} model that Mongan et al. used in their study.[7]

The model compounds have the residue sequence *ACE-X-NME*, where *ACE* is a neutral acetyl capping residue, *X* is the titratable residue, and *NME* is a neutral methyl amine capping residue.[7] The available titratable residues in AMBER are aspartate (AS4), glutamate (GL4), histidine (HIP), lysine (LYS), tyrosine (TYR), and cysteine (CYS), which are all defined as described by Mongan et al.[7] A 10 Å TIP3P[39] solvent buffer was added in a truncated octahedron around each model compound that we simulated. The aspartate model compound was also simulated with larger box sizes—15 Å and 20 Å buffers—to determine if it had any effect on the calculated pK_a.

After the system topologies were generated, each system was minimized using 100 steps of steepest-descent minimization followed by 900 steps of conjugate gradient minimization. They were then heated at constant pressure, varying the target temperature linearly from 50 to 300 K over 200 ps. The solvated model compounds were then run for 2 ns at constant temperature and pressure.

Each model compound system was simulated at constant pH and volume for 2 ns, setting τ_{rlx} = 200*fs*. Each simulation employed pH-REMD using six replicas with the solution pH set to pK_a ± 0.1, pK_a ± 0.2, and pK_a ± 1.2. To evaluate the effect of the solvent relaxation time, the cysteine model compound was run with τ_{rlx} set to 10 fs, 40 fs, 100 fs, 200 fs, and 2 ps.

3.3. ACFCA. A pentapeptide with the sequence *Ala-Cys-Phe-Cys-Ala* (ACFCA) was solvated with a 15 Å buffer of TIP3P molecules around the solute in a truncated octahedron.

The system was minimized using 100 steps of steepest descent minimization followed by 900 steps of conjugate gradient. The minimized structure was heated by varying the target temperature linearly from 50 to 300 K over 200 ps at constant pressure. The resulting structure was then simulated at 300 K at constant temperature and pressure to stabilize the system density and equilibrate the solvent distribution around the small peptide.

The resulting structure was then used in simulations at six different solution pH values—7.1, 8.1, 8.3, 8.7, 8.9, and 9.9. These pH values were chosen because the pK_a of the cysteine model compound is 8.5, so the two cysteines of ACFCA were expected to titrate in this pH range. To demonstrate the effect that pH-REMD had on the titration of ACFCA, two sets of simulations were run—CpHMD with no exchanges and pH-REMD—with each replica being run for 2 ns.

3.4. Hen Egg White Lysozyme. We used the structure solved in PDB code 3LZT as the starting structure for our simulations.[40] All eight aspartate residues were renamed AS4, both glutamate residues were renamed GL4, and histidine 15 was renamed HIP in preparation for the pH-REMD simulations.

All disulfide bonds were added manually in *tleap*, and the system was solvated with a 10 Å TIP3P water buffer surrounding the protein in a truncated octahedron. We added 26 sodium ions and 11 chloride ions in random locations in the

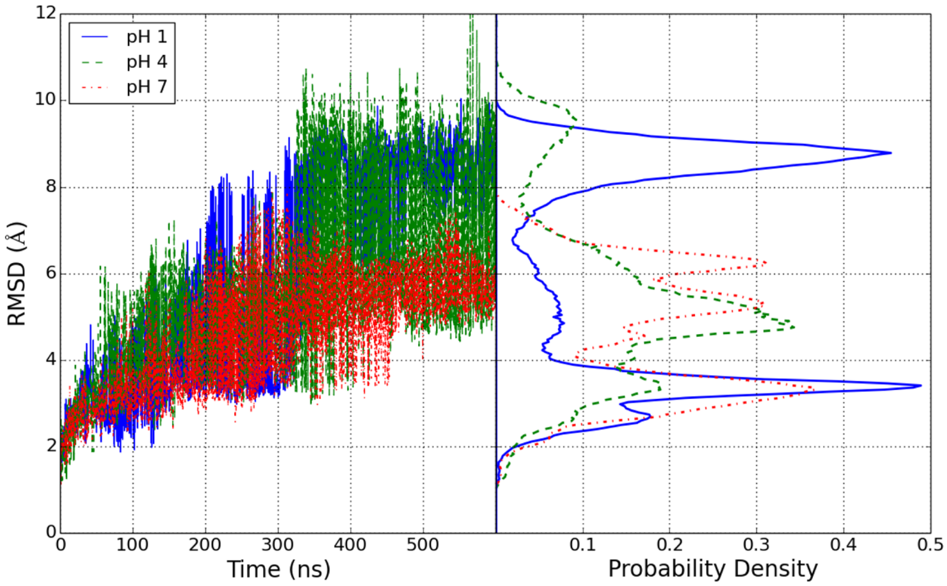


Figure 2. RMSD compared to the 1AKI crystal structure for the ensembles at pH 1, 4, and 7. The time series is shown on the left with the normalized histograms shown on the right.

simulation cell to provide an ionic atmosphere and neutralize the system in its initial, default protonation states.

The system was minimized using 1000 steps of steepest descent minimization followed by 4000 steps of conjugate gradient, with 10 kcal/mol·Å^2 positional restraints applied to the backbone. The structures were then heated at constant volume, varying the target temperature linearly from 10 to 300 K over 400 ps using Langevin dynamics with a 5 ps^{-1} friction coefficient. The heated structures were then equilibrated for 4 ns at constant temperature and pressure using Langevin dynamics with a 2 ps^{-1} friction coefficient to maintain the temperature and 2 ps^{-1} coupling constant for the Berendsen barostat. The system was then subjected to 112 ns of equilibration MD at constant volume and temperature using Langevin dynamics with a 2 ps^{-1} friction coefficient to maintain a constant temperature of 300 K.

Following the setup stages of the simulations, two sets of pH-REMD simulations were carried out for 122 and 150 ns with 12 replicas spanning the pH range from −3 to 8 at 1 pH-unit intervals to characterize the acidic-range titration behavior of HEWL.

3.5. Simulation Details.

All systems were parametrized using the AMBER FF10 force field, which is equivalent to the AMBER FF99SB force field for proteins.[19] The *tleap* program of the AmberTools 12 program suite was used to build the model compound and ACFCA molecules, to add hydrogen atoms HEWL and to solvate each system that was run in explicit solvent.

All simulations were performed using either the *sander* or *pmemd* module of a development version of AMBER 12.[41] Langevin dynamics was used in every simulation to maintain constant temperature with collision frequencies varying from 1 ps^{-1} to 5 ps^{-1}, and the random seed was set from the computer clock to avoid synchronization artifacts.[42,43] The Berendsen barostat was used to maintain constant pressure for the equilibration dynamics with a coupling constant of 1 ps^{-1}.

All molecular dynamics, including the solvent relaxation dynamics, are run with a 2 fs time step, constraining bonds containing hydrogen using SHAKE.[44,45] Replica exchange attempts between adjacent replicas were made every 200 fs for all pH-REMD simulations. Protonation state changes were attempted every 200 fs for all constant pH simulations.

Long-range electrostatic interactions were treated with the particle-mesh Ewald method[46,47] using a direct-space and van der Waals cutoff of 8 Å. Defaults were used for the remaining Ewald parameters. The GB model proposed by Onufriev et al., indicated by the parameter *igb* = 2 in *sander*,[31] was used to evaluate the protonation state change attempts to be consistent with the original implementation in implicit solvent.[7] The intrinsic solvent radius of the carboxylate oxygen atoms was reduced from the standard value of 1.5 Å to 1.3 Å to compensate for the effect of having 2 dummy protons present on each oxygen in the syn- and anti-positions in the explicit solvent simulations.[7]

The deprotonation fraction (f_d) and pH for each simulation—and each window of the running averages—was fitted to the Hill equation (eq 4) using the Levenberg–Marquardt nonlinear least-squares algorithm implemented in *SciPy* to compute the pK_a and Hill coefficient (n). All pK_a values reported for titratable residues in this paper correspond to the value computed by fitting f_d from the simulations at every pH to eq 4 over the specified time interval.

$$f_d = \frac{1}{1 + 10^{n(\mathrm{p}K_a - \mathrm{pH})}} \tag{4}$$

4. RESULTS AND DISCUSSION

First, we will analyze the long simulations in implicit solvent that demonstrate the weaknesses of implicit solvent models. Next, we will analyze our proposed CpHMD and pH-REMD methods as well as ways to optimize its overall performance. We will start by discussing the behavior of the model compounds when the size of the unit cell and the length of the relaxation dynamics (τ_{rlx}) is varied. We will follow this discussion with a similar analysis on a slightly larger system—ACFCA—before discussing the application of our proposed method to HEWL.

4.1. Long GB Simulations. Long simulations are required to observe rare events or phenomena that occur on long time scales, but often expose deficiencies in a computational model that are frequently missed in shorter simulations.[48] Therefore, we ran CpHMD GB simulations of HEWL for 600 ns—more than an order of magnitude longer than what has previously been published—to determine if these calculations were stable and could be used to study the dynamical behavior of proteins over long time scales.

The backbone RMSD with respect to the starting crystal structure, shown in Figure 2 for the replicas at pH 1, 4, and 7, reaches as much as 10 Å at lower pH values, suggesting that the native state is unstable in GB implicit solvent at long time scales. As the lysozyme is active at the lysosomal pH around 4.5, CpHMD simulations near this pH should reflect this stability. As the backbone RMSD from the crystal structure increases, the predicted pK_a value worsens. We computed the root mean squared error (RMSE) of the computed acidic residue pK_a values compared to experiment. The predicted pK_a was computed for each residue by taking a running average with a 10 ns window of the deprotonation fraction for each residue at each pH in the pH-REMD simulation. The RMSE of the calculated pK_a values from experiment averaged over all titratable residues is shown in Figure 3.

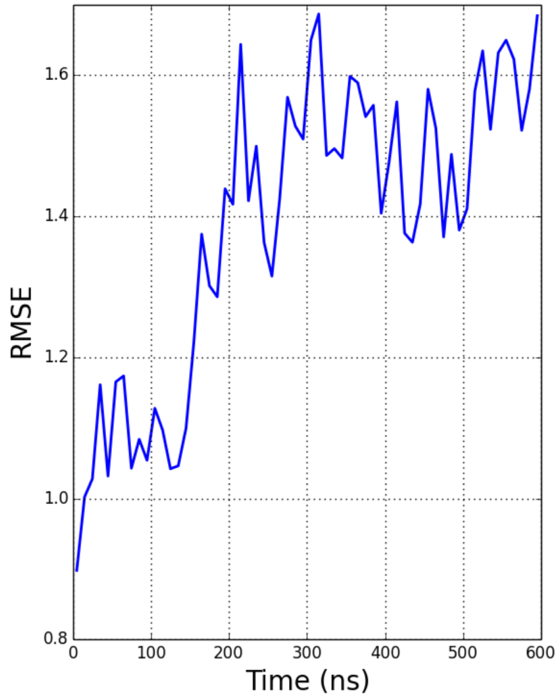


Figure 3. RMSE of all acidic titratable residue pK_a values compared to experiment[49] during the course of the simulation. A 10 ns window was used for the running average of the computed pK_a.

4.2. Explicit Solvent Simulations. *4.2.1. Box Size Effects.* To study the effect that the unit cell size has on titrations in our proposed method, we prepared three simulations of the aspartate model compound with different TIP3P solvent buffers surrounding it. We prepared systems with a 10 Å, 15 Å, and 20 Å TIP3P solvent buffer around the model aspartate.

Because protonation state sampling takes place in GB solvent without periodic boundary conditions, any effect of the box size on calculated pK_a values can only arise due to alterations of the structural ensembles induced by artifacts from the box size. The calculated pK_a values of the three systems were 4.02 ± 0.07, 4.05 ± 0.08, and 4.12 ± 0.07 for the 10 Å, 15 Å, and 20 Å solvent buffer systems, respectively. To estimate the uncertainties we divided each simulation into 100 ps chunks and took the standard deviation of the set of 20 pK_a values calculated from those segments.

To further demonstrate the insensitivity of box size to pH-REMD titrations, we plotted the solvent radial distribution functions (RDFs) around the center of mass of the carboxylate functional group in Figure 4. The insensitivity of the pK_a and solvent structure with respect to the model compound provides strong evidence that no undue care is necessary when choosing the size of the solvent buffer for these types of simulations.

4.2.2. Effect of Solvent Relaxation Time (τ_{rlx}). An important approximation in the proposed method is that the protonation state sampling ρ' from eq 2 can be replaced using an implicit solvent model followed by relaxation MD to generate the relaxed solvent positions and momenta. The question then becomes how long this relaxation dynamics should be run.

To address this, 2 ns of constant protonation molecular dynamics simulations were run on the model cysteine compound in both protonation states—protonated and deprotonated—after the same minimization and heating protocols were used as for the other model compound simulations. The protonation state was then swapped for the final structures of both simulations, and MD was performed while constraining the solute position for 20 ns, equivalent to the relaxation dynamics protocol in our proposed method.

The optimum value for τ_{rlx} is the time after which the energy of the relaxation trajectory stabilizes and the simulation loses all memory of its initial configuration. To be truly equivalent to having been chosen at random, the final, relaxed solvent distribution must be completely uncorrelated from the initial distribution at the time the protonation state was changed.

To probe the necessary time scales for these relaxation dynamics, the energy of each snapshot in the relaxation trajectory is plotted alongside the autocorrelation function of that energy in Figure 5 to clearly demonstrate the 'appropriate' value of τ_{rlx} for this model system.

We chose the cysteine model compound for this test for two reasons. First, the model compounds are fully solvent-exposed due to their small size, which results in a worst-case scenario in terms of the number of water molecules that must be reorganized during the relaxation dynamics. The optimum τ_{rlx} value for model compounds is expected to be an upper-bound on the values required for larger systems. Second, cysteine is the smallest and simplest of the titratable amino acids, eliminating potential complications from tautomeric states compared to aspartate, glutamate, and histidine.

The relaxation energies plotted in Figure 5 begin to stabilize after 4 to 6 ps of relaxation dynamics and are largely uncorrelated within that same time scale. However, because 4 ps of MD—corresponding to 2000 steps of dynamics with a 2 fs time step—adds dramatically to the cost of CpHMD simulations in explicit solvent, we will analyze the approximation of using a significantly smaller value for τ_{rlx}.

Both the relaxation energies and autocorrelations drop very sharply at the start of the relaxation dynamics, so the majority of the benefit gained by relaxing the solvent is realized within the first few steps.

For both simulations, the first 200 fs of relaxation dynamics resulted in 70% of the total relaxation energy in calculations.

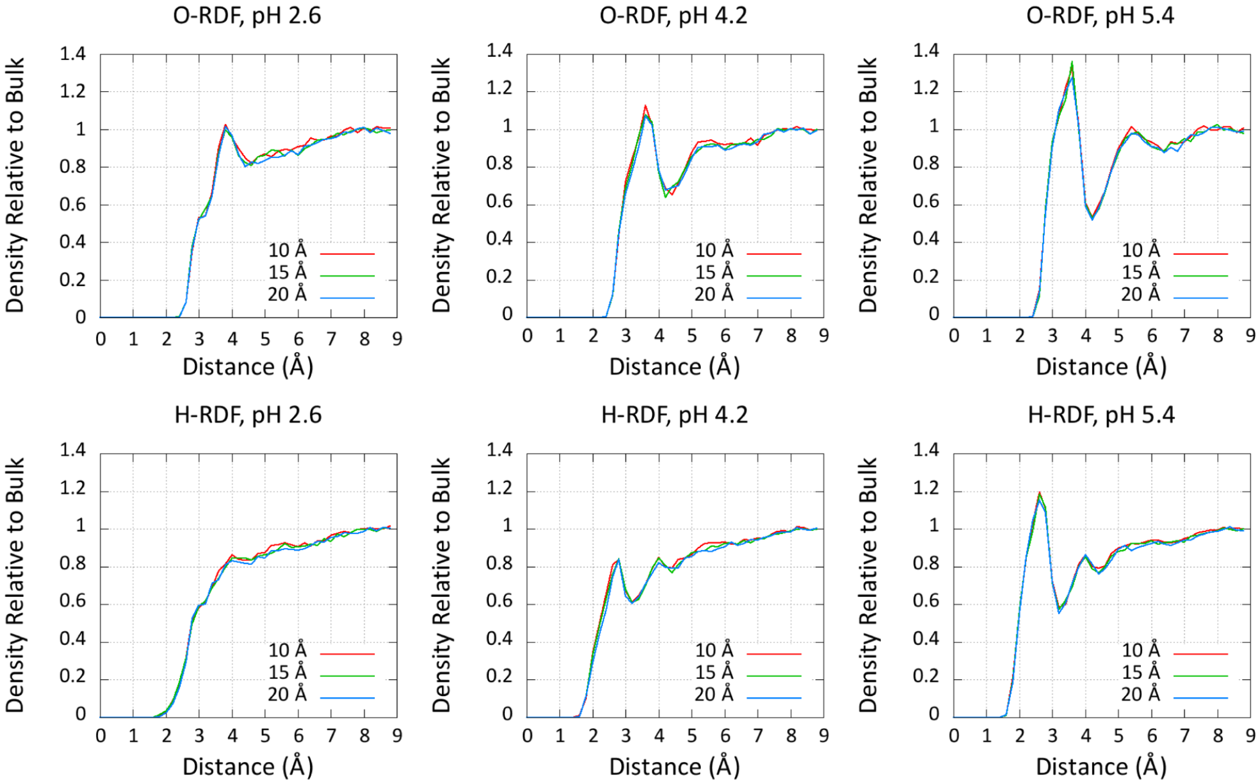


Figure 4. Radial distribution functions (RDFs) of solvent oxygen atoms (O) and hydrogen atoms (H) with different unit cell sizes. The shown measurements—10, 15, and 20 Å—represent the size of the solvent buffer surrounding the solute. RDF plots for three different pHs are shown, highlighting the pH dependence of the solvent structure around the carboxylate of the aspartate model compound and its invariance to box size.

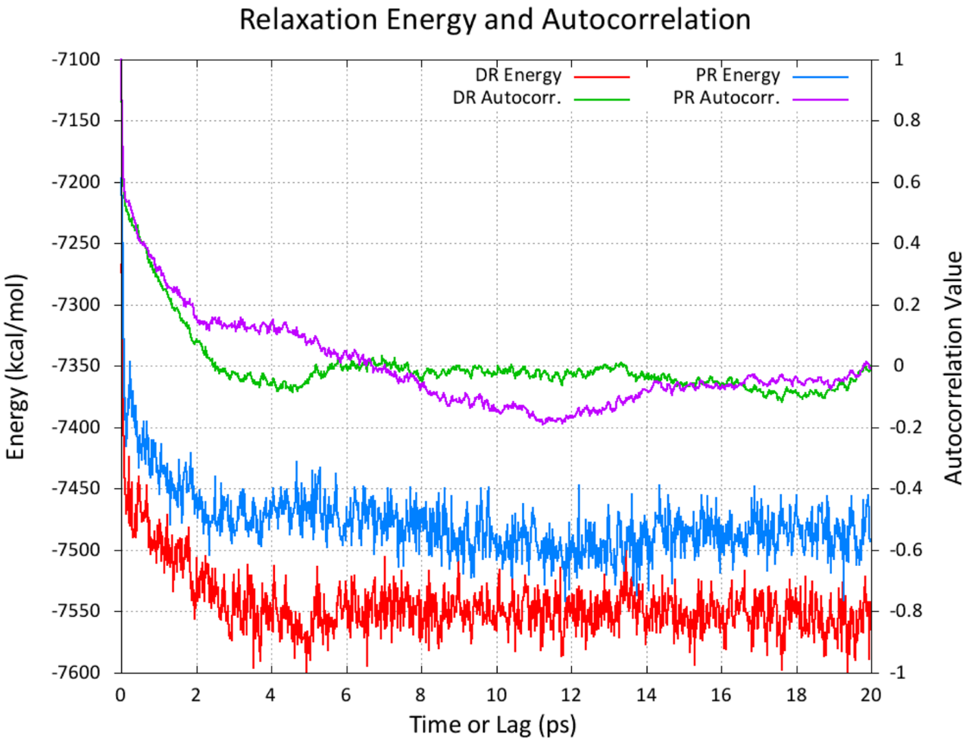


Figure 5. The relaxation of the protonated state, starting from the protonated trajectory, is shown in blue with its autocorrelation function shown in purple. The relaxation of the deprotonated state from an equilibrated snapshot from the protonated ensemble is shown in red with its autocorrelation function shown in green. Here, PR and DR stand for *Protonated-Relaxation* and *Deprotonated-Relaxation*, respectively.

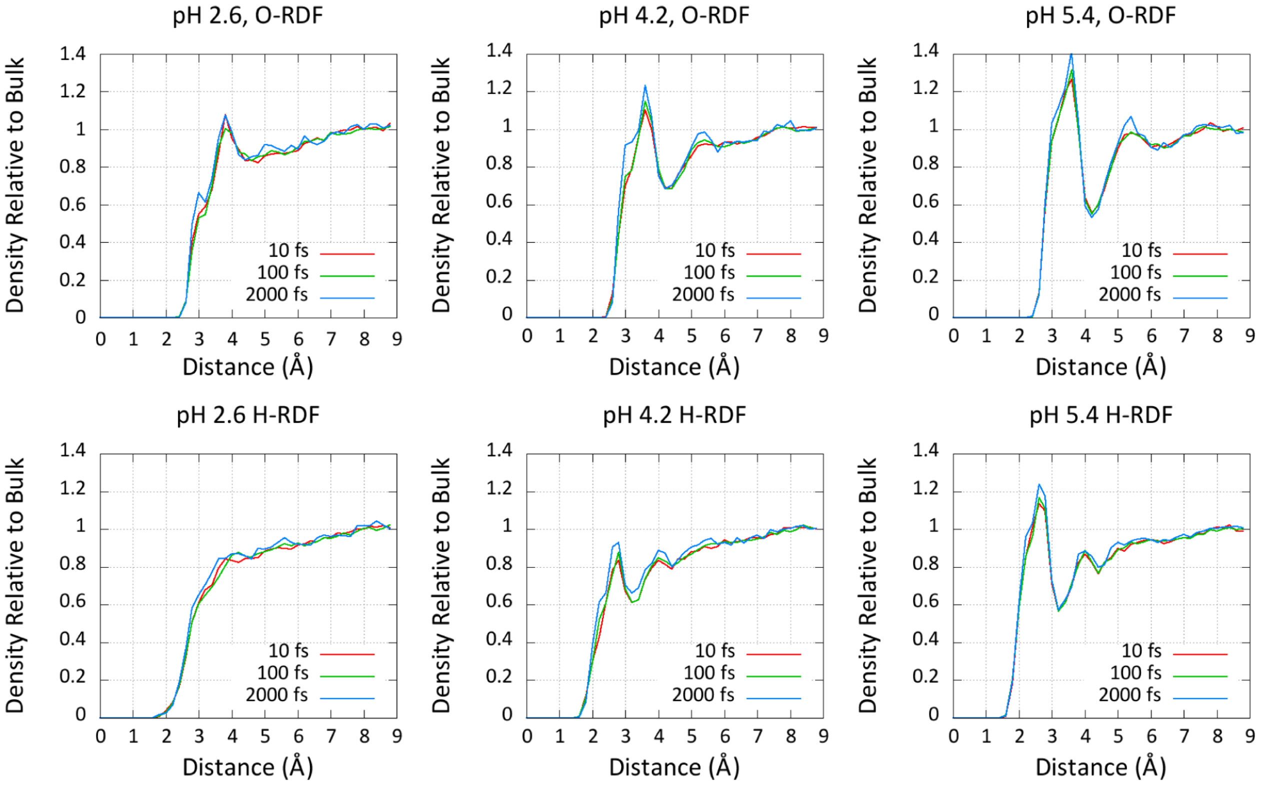


Figure 6. RDFs of water oxygen atoms (O) and hydrogen atoms (H) around the center-of-mass of the carboxylate group of the model aspartate molecule at different solution pHs.

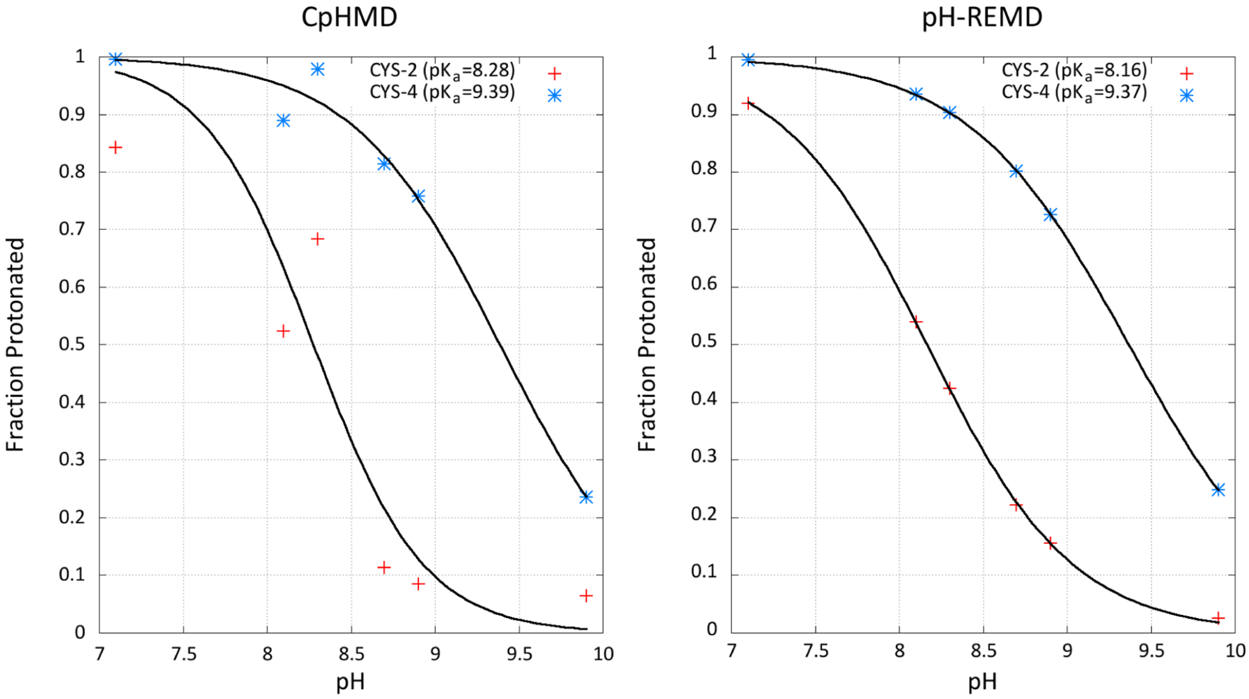


Figure 7. Titration curves of Cys-2 and Cys-4 in the *ACFCA* pentapeptide. Results from CpHMD (no replica exchange attempts) and pH-REMD are shown in the plots on the left and right, respectively.

The correlation of the relaxation energy decays similarly, so the assumption that the relaxed solvent distribution is uncorrelated from its starting point is a reasonable approximation.

To validate the use of a shorter τ_{rlx}, we titrated the aspartate model compound using pH-REMD with five different values for τ_{rlx}—10 fs, 40 fs, 100 fs, 200 fs, and 2 ps. The calculated pK_a for each simulation was 4.10, 4.08, 4.07, 4.10, and 4.05, respectively, for the listed relaxation times. Furthermore, comparing the solvent radial distribution functions of the different solvent relaxation times (Figure 6) shows little dependence of the solvent distribution on the value of τ_{rlx}.

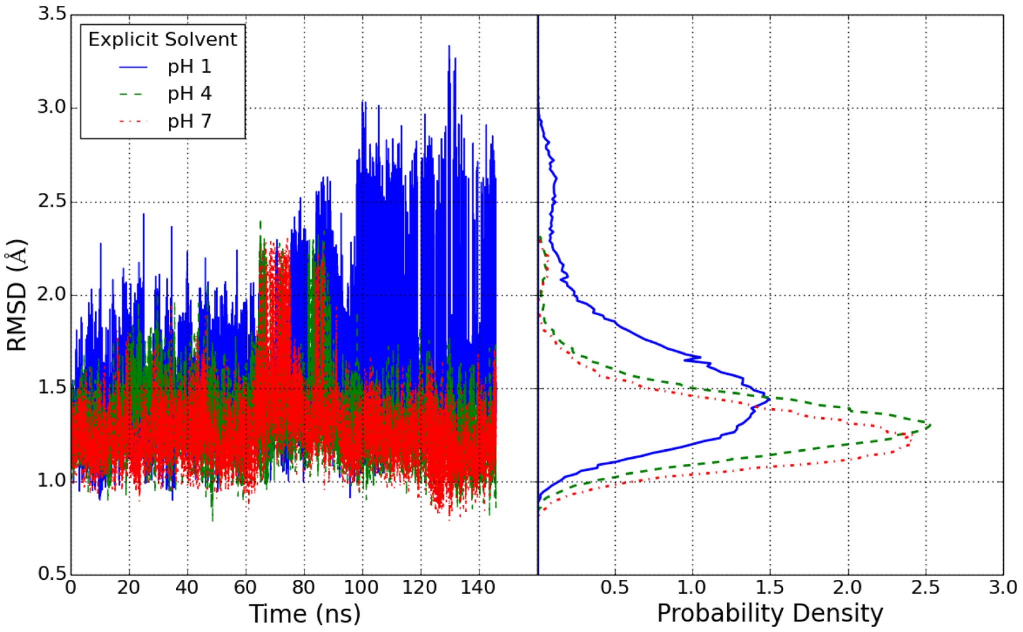


Figure 8. Backbone RMSD distributions for HEWL simulations in explicit solvent with solution pH set to 1, 4, and 7. Time series are shown on the left and histograms are shown on the right.

4.2.3. ACFCA: CpHMD vs pH-REMD. The small peptide chain ACFCA, described in section 3.3, was chosen as a test due to its small size and predictable titration behavior. The simplicity of the system makes it an ideal test; its small size mitigates the conformational sampling problem, and the simple titrating behavior of cysteine further simplifies protonation state sampling. Unlike aspartate and glutamate, which have the four defined tautomeric states defined by anti- and syn-protonation on each of two carboxylate oxygens, and histidine which has two tautomeric states on the imidazole, cysteine has only one protonated and one deprotonated state, presenting fewer degrees of freedom that must be exhaustively sampled.

Each cysteine is in a slightly different microenvironment due to the different charges of the N- and C-termini. Due to its proximity to the N-terminus, Cys-2 is expected to experience a negative pK_a shift with respect to the model compound due to the electrostatic interactions with the positively charged terminus. Cys-4, on the other hand, is expected to experience a pK_a shift in the opposite direction due to the electrostatic interactions with the negatively charged C-terminus.

We ran simulations at pH 7.1, 8.1, 8.3, 8.7, 8.9, and 9.9 to sufficiently characterize the titration behavior of both cysteine residues around their pK_a values. One set of replicas was run with pH-REMD while the other set was run using CpHMD (i.e., without attempting exchanges between the replicas). The titration curves for both sets of simulations, shown in Figure 7, demonstrate the importance of using pH-REMD in constant pH simulations in explicit solvent. As expected, the pH-REMD simulations revealed pK_a shifts of −0.2 pK units for Cys 2 and +0.9 pK units for Cys 4 with respect to the model Cys compound.

Even for a simple system such as ACFCA, using pH-REMD on top of standard CpHMD simulations results in a drastic improvement in titration curve fit—a result of improved protonation state sampling. The residual sum of squares (RSS), a quantity that measures how well an equation fits a data set, shows drastic improvement using pH-REMD. The RSS for Cys-2 and Cys-4 using CpHMD was 9×10^{-2} and 7×10^{-3}, respectively. For the pH-REMD simulations, on the other hand, the RSS was reduced by several orders of magnitude to 7×10^{-5} and 9×10^{-6} for Cys-2 and Cys-4, respectively.

4.2.4. Hen Egg White Lysozyme. HEWL is a common benchmark for pK_a calculations because it has been studied extensively both experimentally[49−51] and theoretically,[7,9,24,26,34,52] and it has a large number of titratable residues—some with a marked pK_a shift compared to the isolated model compound.

The simulations run in explicit solvent revealed far more stable trajectories over all pH values than their analogs run in implicit solvent over the 150 ns time scale of the explicit solvent simulations. In Figure 8, the plots of backbone RMSDs are bounded below 3.5 Å for the duration of the ca. 150 ns simulation. The RMSD distributions from the second 120 ns simulation are very similar.

The predicted pK_a values for the titratable residues, summarized in Table 1 for both sets of simulations, show good agreement to experiment. The agreement is significantly better than the predictions from the implicit solvent calculations over a similar time scale. With the exception of aspartate 119 (and aspartate 52 in the 150 ns simulation), the predicted pK_a values of all residues were within 1 pK unit from the experimental values given in ref 49.

Furthermore, the large fluctuations in the RMSE throughout the course of the simulation—between 0.70 and 1.25 seen in Figure 9—suggest that differences in reported RMSE around 0.1 pK units are statistically insignificant for short CpHMD and pH-REMD simulations on the order of 10 ns. The standard deviation of the pK_a RMSE plotted in Figure 9 is 0.12 pK units with a correlation time of roughly 25 ns. The standard error of the mean, given by $(\sigma/N)^{1/2}$ where σ is the variance and N is the number of independent samples, is 0.05 pK units over 150 ns. Therefore, there is no statistically significant difference between methods with reported pK_a RMSEs within 0.10 pK units of each other (±0.05 pK units for each simulation) over 150 ns.

Table 1. Calculated pK_a Values for Acid-Range Titratable Residues in HEWL Using the Proposed Method for a ca. 150 ns Simulation and a ca. 120 ns Simulation[a]

		simulation		
residue	150 ns	120 ns	implicit	experiment
Glu 7	3.37	3.31	3.85	2.6
His 15	6.38	6.32	5.79	5.5
Asp 18	2.83	2.89	2.49	2.8
Glu 35	6.27	6.32	3.65	6.1
Asp 48	2.31	1.92	2.47	1.4
Asp 52	2.24	2.63	3.37	3.6
Asp 66	1.87	1.81	1.50	1.2
Asp 87	2.02	2.06	2.98	2.2
Asp 101	4.36	4.25	2.32	4.5
Asp 119	1.53	1.53	1.74	3.5
RMSE	0.92	0.82	1.32	

[a] Experimental values are taken from ref 49. Implicit solvent results are shown for comparison. (Results are shown for the final 500 ns of the 600 ns implicit solvent simulation).

Unlike the implicit solvent calculations, the quality of the predicted pK_a values in the proposed method does not appear to degrade over time in long simulations. The deviation of each residue compared to experiment and the total RMSE is shown in Figure 9. The pK_a at each time step is computed from a running average of deprotonation fraction for replicas at each pH using a window size of 5 ns and fitting to eq 4 the same way as described for Figure 3.

The smaller backbone RMSD and drastically improved pK_a predictions are strong evidence that the current method improves conformational sampling by employing an explicit representation of the solvent. Because the protonation state sampling is performed with the same GB potential in both the original and proposed methods, any differences in the predicted pK_a values must be caused by differences in conformational sampling. Given the rigorous nature of the experimental measurements,[49] this provides strong evidence that the proposed method improves significantly upon the original by improving conformational sampling with an explicit solvent representation.

5. CONCLUSION

We have extended the constant pH molecular dynamics method developed by Mongan et al.[7] so that the dynamics can be run in explicit solvent. We tested a wide range of parameters in our proposed method for their effect on the conformational and protonation state sampling of small test systems. Because these test systems are small and their titratable sites are completely solvent-exposed, they likely represent the highest level of sensitivity to these various parameters.

In particular, we found that the box size of the unit cell had no discernible effect on the titration behavior of the aspartate model compound, given cell sizes that ranged from 20 Å in diameter—one of the smallest sizes permissible when using the minimum image convention with an 8 Å cutoff—to 40 Å in diameter.

Another key aspect of the current method is the necessity to relax the solvent around any new protonation state selected by the MC moves carried out in GB. By analyzing the decay of the potential energy in the solvent relaxation dynamics, we determined that 4 ps of MD was sufficient to stabilize the energy of the solvent distributions and generate relaxed solvent

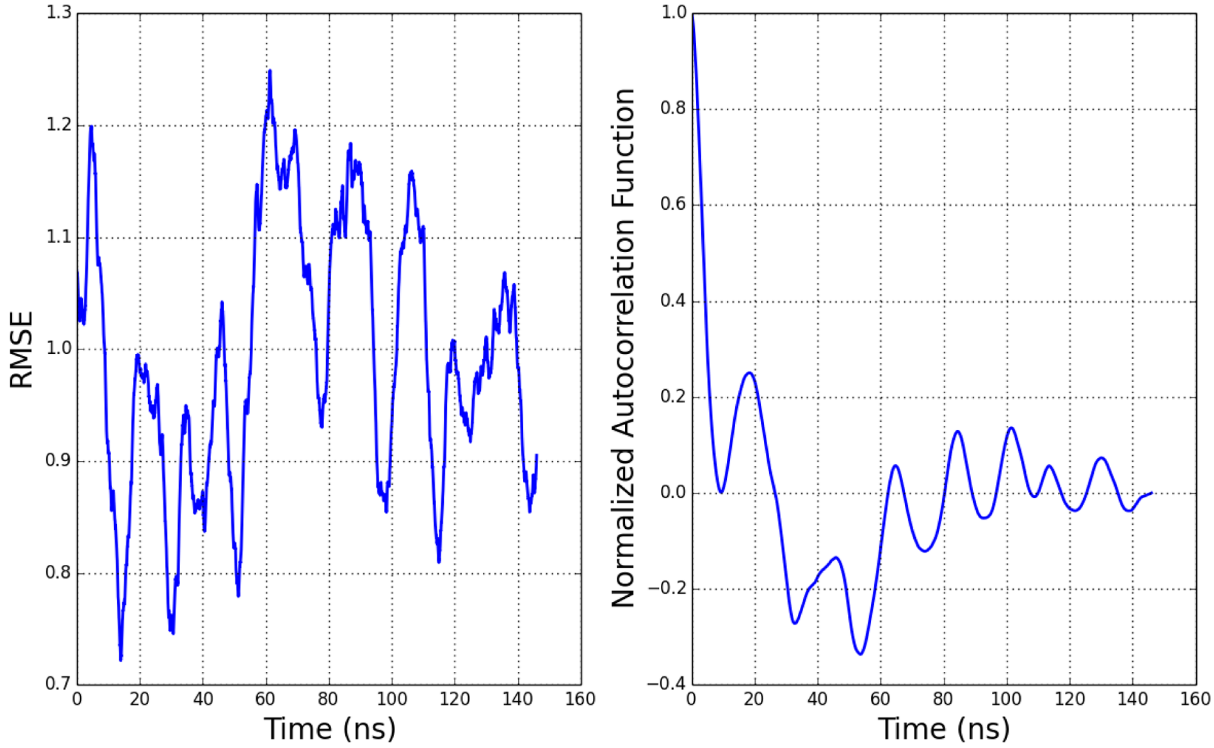


Figure 9. (left) Plot showing the RMSE of all acidic titratable residue pK_a values compared to experiment[49] during the course of the simulation. A 5 ns window was used for the running average of the computed pK_a. (right) Plot showing the autocorrelation function of the RMSE time series, showing that the correlation time between statistically independent samples is roughly 25 ns.

conformations whose energies are uncorrelated from the initial arrangements. However, given the expense of such a long relaxation period, we investigated using fewer relaxation steps to increase the simulation efficiency and found shorter times—down to 0.2 ps—had no measurable effect on the calculated pK_a and very little effect on the solvent distribution around the model cysteine compound.

Further tests on a small pentapeptide test system with two titratable sites (ACFCA) showed the importance of using pH-REMD over conventional CpHMD with the proposed method. While we showed that the enhanced protonation state sampling of pH-REMD results in smoother titration curves for complex proteins in implicit solvent,[24] even the simplest systems in explicit solvent require pH-REMD to obtain a smooth titration curve.

We tested the proposed method on HEWL, a very common pK_a benchmark system. We found that the proposed method of using GB implicit solvent to evaluate protonation state changes and explicit solvent to propagate dynamics yielded stable trajectories whose predicted pK_a values agreed well with experiment. Our results show that using the proposed method leads to a significant improvement in how systems are modeled at constant pH compared to the original method that used GB to propagate system dynamics.

Often, the most interesting titratable residues in biological systems have a large pK_a shift compared to the model compound. These highly perturbed residues have environments drastically different than the one provided by bulk solvent, and the conformational sampling must be both accurate and extensive to yield accurate pK_a predictions. In future work we will explore the use of enhanced sampling techniques in conjunction with pH-REMD in an attempt to improve the efficiency of the conformational sampling in explicit solvent, such as accelerated MD and temperature-based REMD.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: roitberg@ufl.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.M.S. gratefully acknowledges support from the National Science Foundation (NSF) GRFP award, computational support from NCSA BlueWaters and the University of Florida High Performance Computing center, and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant No. OCI-1053575. The authors acknowledge support from NSF ACI-1147910 and ACI-1036208 to A.E.R. and National Institutes of Health (NIH) GM62248 to D.M.Y.

■ REFERENCES

(1) Garcia-Moreno, B. *J. Biol.* **2009**, *8*, 98.

(2) Perutz, M. F. *Science* **1978**, *201*, 1187−1191.

(3) Alexov, E.; Mehler, E. L.; Baker, N.; Huang, Y.; Milletti, F.; Nielsen, J. E.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. *Proteins* **2011**, *79*, 3260−3275.

(4) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins* **1997**, *27*, 523−544.

(5) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184−4200.

(6) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III *Proteins* **2004**, *56*, 738−752.

(7) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038−2048.

(8) Khandogin, J.; Brooks, C. L., III *Biophys. J.* **2005**, *89*, 141−157.

(9) Wallace, J. A.; Shen, J. K. *J. Chem. Theory Comput.* **2011**, *7*, 2617−2629.

(10) Goh, G. B.; Knight, J. L.; Brooks, C. L. *J. Chem. Theory Comput.* **2012**, *8*, 36−46.

(11) Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmüller, H. *J. Chem. Theory Comput.* **2011**, *7*, 1962−1978.

(12) Itoh, S. G.; Damjanović, A.; Brooks, B. R. *Proteins* **2011**, *79*, 3420−3436.

(13) Walczak, A. M.; Antosiewicz, J. M. *Phys. Rev. E* **2002**, *66*, 051911.

(14) Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins* **2002**, *47*, 469−480.

(15) Machuqueiro, M.; Baptista, A. M. *Proteins* **2011**, *79*, 3437−3447.

(16) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E., III; JureČka, P. *J. Chem. Theory Comput.* **2011**, *7*, 2886−2902.

(17) Cheatham, T. E., III; A.Young, M. *Biopolymers* **2001**, *56*, 232−256.

(18) Várnai, P.; Djuranovic, D.; Lavery, R.; Hartmann, B. *Nucleic Acids Res.* **2002**, *30*, 5398−5406.

(19) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−725.

(20) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120−127.

(21) Srivastava, J.; Barber, D. L.; Jacobson, M. P. *Physiology* **2007**, *22*, 30−39.

(22) Frantz, C.; Barreiro, G.; Dominguez, L.; Xiaoming, C.; Eddy, R.; Condeelis, J.; Kelly, M. J. S.; Jacobson, M. P.; Barber, D. L. *J. Cell Biol.* **2008**, *183*, 865−871.

(23) Di Russo, N.; Estrin, D. A.; Martí, M. A.; Roitberg, A. E. *PLoS Comput. Biol.* **2012**, *8*, 1−9.

(24) Swails, J. M.; Roitberg, A. E. *J. Chem. Theory Comput.* **2012**, *8*, 4393−4404.

(25) Wallace, J. A.; Shen, J. K. *J. Chem. Phys.* **2012**, *137*, 184105.

(26) Machuqueiro, M.; Baptista, A. M. *Proteins* **2008**, *72*, 289−298.

(27) Baptista, A. M.; Soares, C. M. *J. Phys. Chem. B* **2001**, *105*, 293−309.

(28) Manousiouthakis, V. I.; Deem, M. W. *J. Chem. Phys.* **1999**, *110*, 2753−2756.

(29) Gregory D., Hawkins; C., C.; Truhlar, D. *Chem. Phys. Lett.* **1995**, *246*, 122−129.

(30) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(31) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383−394.

(32) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156−169.

(33) Shang, Y.; Nguyen, H.; Wickstrom, L.; Okur, A.; Simmerling, C. *J. Mol. Graphics* **2011**, *29*, 676−684.

(34) Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 560−568.

(35) Bogusz, S.; Cheatham, T. E., III; Brooks, B. R. *J. Chem. Phys.* **1998**, *108*, 7070−7084.

(36) Rocklin, G. J.; Mobley, D. L.; Dill, K. A.; Hünenberger, P. H. *J. Chem. Phys.* **2013**, *139*, 184103.

(37) Götz, A.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput.* **2012**, *8*, 1542.

(38) Artymiuk, P. J.; Blake, C. C. F.; W., R. D.; S., W. K. *Acta Cryst. B* **1982**, *38*, 778−783.

(39) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(40) Walsh, M. A.; Schneider, T. R.; Sieker, L. C.; Dauter, Z.; Lamzin, V. S.; Wilson, K. S. *Acta Crystallogr. D* **1998**, *54*, 522−546.

(41) Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Götz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, C., Sagui, R.; ; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*; University of California: San Francisco, 2012.

(42) Uberuaga, B. P.; Anghel, M.; Voter, A. F. *J. Chem. Phys.* **2004**, *120*, 6363−6374.

(43) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. *J. Chem. Theory Comput.* **2009**, *5*, 1624−1631.

(44) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(45) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952−962.

(46) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(47) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Hsing, L.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(48) Cheatham, T. E., III *Curr. Opin. Struct. Biol.* **2004**, *14*, 360−367.

(49) Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O'Meara, F.; Sondergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.; Nielsen, J. E. *Proteins* **2011**, *79*, 685−702.

(50) Takahashi, T.; Nakamura, H.; Wada, A. *Biopolymers* **1992**, *32*, 897−909.

(51) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180−1184.

(52) Demchuk, E.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 17373−17387.