Titr-DMD – A Rapid Constant pH Molecular Dynamics Framework

David J. Reilley,† Jian Wang,* Nikolay V. Dokholyan,‡§ Anastassia N. Alexandrova*†

† Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90095-1569, United States
‡ Department of Pharmacology, Department of Biochemistry & Molecular Biology, Penn State University College of Medicine, Hershey, Pennsylvania 17033, United States
§ Department of Chemistry and Biomedical Engineering, Pennsylvania State University, University Park, Pennsylvania 16802, United States
∥ California NanoSystems Institute, Los Angeles, California 90095-1569, United States

KEYWORDS: CpHMD, pK_a, DMD, Dynamics

ABSTRACT: The pH dependence of enzyme fold stability and catalytic activity is a fundamentally dynamic, structural property which is difficult to study. Computational methods, particularly constant pH molecular dynamics (CpHMD), are the best situated tools for this. However, these often struggle with affordable sampling of sufficiently long timescales, accuracy of pK_a prediction, and verification of the structures they generate. We introduce Titr-DMD, an affordable CpHMD method with a protonation state sampler that can be systematically improved, to circumvent these issues. We benchmark the method on a set of proteins with experimentally attested pK_a and on the pH triggered conformational change in a staphylococcal nuclease mutant, a rare experimental study of such behavior. Our results show Titr-DMD to be an effective method to study pH coupled protein dynamics.

Introduction

Solution pH is a chemical property with an immense effect on protein behaviors that are difficult to study at the atomic scale. Peak protein fold stability and catalytic activity are both dependent on an often narrow range of pH. Understanding the sequential and structural underpinning of these preferences contributes to the design and application of enzymes, particularly extremophile enzymes – which would allow for their use in harsher reaction conditions in industrial catalysis.¹–⁴ and answers a wide range of questions of medical interest as precise pH regulation is critical for cellular homeostasis.⁵–⁷ However, this understanding demands atomistic information of fundamentally dynamic phenomena. pH-dependent dynamics is challenging to study experimentally, requiring a combination of techniques like NMR monitored pH-titration, circular dichroism spectroscopy, and X-ray crystallography none of which alone provide the complete picture. This experimental complexity leaves computational investigation⁸ as a critical tool to fill in the gaps.

Successful computational methods that assess pH dependent protein behavior must accurately couple amino acid protonation state change with conformational dynamics. Typically, continuum electrostatic methods describe the protonation states of amino acids, assessing the free energy of protonation and deprotonation events or pK_a. Various solutions to the Poisson-Boltzmann equation can provide this,⁹ especially the generalized Born model.¹⁰,¹¹ Tools such as UHBD,¹² H⁺⁺,¹³ and Propka¹⁴,¹⁵ predict the pK_a of amino acid residues in proteins based on static structures, but can not fully capture pH-dependent dynamic behavior. Molecular dynamics can provide the missing conformational sampling. The combinations of these tools are known as constant pH molecular dynamics (CpHMD); these methods generally use continuum electrostatic methods to model the protonation state changes of amino acids over the course of a molecular dynamics simulation.

The appropriate sampling of pH coupled dynamics is difficult to achieve for all CpHMD methods and challenging to verify. The choice of solvation model is central to sampling and broadly breaks CpHMD methods into two categories. Implicit solvent-based methods offer rapid sampling with the treatment of the surrounding solution as a simple dielectric medium.¹⁶–¹⁹ Explicit solvent-based methods can provide greater accuracy through atomistic solvent treatment, but sufficient sampling is difficult to achieve, as both conformational and protonation states need to be sampled. Furthermore, protonation sampling is affected by poor overlap between solvent configurations such that protonation state changes are often immediately rejected. To counter this, many groups have applied λ-dynamics, based off pioneering work by Brooks et al.²⁰ (in turn based on earlier work with other thermodynamic properties in mind).²¹,²² which treats the protonation state of individual amino acid sites as continuous degrees of freedom rather than discrete ones sampled distinctly.²³–²⁵ Other efforts focus on enhancing/accelerating conformational sampling through GPU processing²⁶ or replica exchange.²⁷,²⁸ Ultimately,
regardless of solvation, one struggle for all CpHMD methods is verification of the ensemble of conformational and protonation states they generate, due to paucity of complementary experimental results. Most methods instead benchmark through more plentiful indirect evidence, such as reconstruction of titration curves or estimation of experimental pKₐ values. It is unclear if the generated ensemble of structures is physically meaningful by comparison to those metrics alone.

We present here Titr-DMD as an undemanding method for the investigation of pH dependent protein behavior. Our method dynamically updates the protonation states of a discrete molecular dynamics (DMD) simulation²⁹-³¹ using pKₐ values generated through the semi-empirical continuum electrostatics method Propka, but is not restricted to this tool. DMD’s square-well potentials (reducing the number of necessary calculations) and implicit solvation provide rapid conformational sampling on limited resources, while a Monte Carlo algorithm confers extensive protonation state sampling. Our program is highly modular for easy modification as better approaches for instantaneous pKₐ prediction develop. We benchmark Titr-DMD on both its ability to calculate ensemble pKₐ compared to experiment and on its ability to capture the pH dependent conformational change found experimentally in staphylococcal nuclease (SNase).³² While other CpHMD methods have studied pH-coupled protein dynamics,³³,³⁴ including in the SNase system,³⁵ recapturing protein dynamics is not a common benchmark for method development.

**Theory and Methods**

**Titr-DMD method.** This method combines rapid DMD conformational sampling with a custom algorithm to resolve protonation based on Propka³¹ pKₐ predictions. Simulations are performed iteratively, alternating between a short DMD simulation and a titration (Titr)-feature that discretely assigns protonation states. The algorithm for the Titr-feature itself comprises five steps: (1) titratable residues are identified, (2) contact networks are constructed from these residues, (3) the solvent accessibility of each network is determined, (4) the probability of protonation state change is determined for each network or residue, (5) protonation state changes are determined by a Monte Carlo scheme (Figure 1).

The intervals between protonation state reassessment are run just long enough so that protonation and deprotonation are equilibrated over the DMD simulation timescale. As isolated proton transfer events, like many individual reaction steps in proteins,³⁶ generally occur on the femtosecond to picosecond timescale,³⁷ 200 DMD steps (which is ~10 ps) suffices, a comfortable separation of 1-3 orders of magnitude. This allows for both the consistent and meaningful application of theory, discussed more thoroughly throughout this section, and extensive sampling of a system’s potential protonation states. A higher reassessment frequency is therefore unnecessary and computationally expensive; while additional time spent on the Titr-feature itself is minimal, a higher frequency requires more, shorter DMD simulations and thus more time overhead during the program initialization.

Selection of titratable residues is based on their solution pKₐ values. The amino acids aspartate, glutamate, histidine, cysteine, tyrosine, lysine, and arginine are the only ones to have side-chain solution pKₐ values in the physiological range of pH 1-13 and so are the only ones considered. While significant shifts in pKₐ often occur when an amino acid is part of a protein, all other residues have side-chain pKₐ that fall far enough out of this range to be largely irrelevant in the vast majority of systems. For the same reason, only the first protonation/deprotonation event is considered for the included amino acids; states like doubly deprotonated lysine or doubly protonated glutamine are inaccessible. The C-terminal carboxylate and N-terminal amine could be titrated as well, but are not currently implemented due to missing DMD potentials for their less preferred states.

Contact networks are constructed on the basis of the proximity of titratable residues. First, interacting pairs of residues are identified based on their (de)protonatable heteroatoms that are within a certain cutoff distance, rₛ, of each other. This pro-

---

**Figure 1: Schematic of the Titr-feature algorithm. This runs between short DMD (or any molecular mechanics) simulations to assign discrete protonation states.**
tonation contact distance \( r_p \) is selected as 3.5 Å to be consistent with the DMD definition of a long hydrogen bond. Each thus defined network represents a series of residues close enough that in the timeframe of the DMD phase of the Titr-DMD simulation the proton exchange is equilibrated between them and lies firmly under thermodynamic control.

Solvent accessibility of each residue contact network is determined in a manner consistent with Propka, which defines a specific residue as buried or exposed based on its contact number, \( w(N) \). \( w(N) \) is determined by the number of heavy atoms, \( N \), within 15 Å of the residue’s charge center according to

\[
w(N) = \begin{cases} 
0 & \text{if } N \leq N_{\text{min}} \\
\frac{N - N_{\text{min}}}{N_{\text{max}} - N_{\text{min}}} & \text{if } N_{\text{min}} < N < N_{\text{max}} \\
1 & \text{if } N \geq N_{\text{max}}
\end{cases}
\]

The residue is thus 0% buried if \( N \leq 280 \) (\( N_{\text{min}} \)) and 100% buried if \( N \geq 560 \) (\( N_{\text{max}} \)). In the Titr-feature, a network is considered solvent accessible if any residue in it is below a certain cutoff. As proton exchange is equilibrated within a network, so long as one residue is solvent accessible the rest of the network can freely exchange protons with solvent. The best value of this solvent access cutoff is a parameter in the model and in general might be system dependent. We find that the most appropriate value for the solvent access cutoff could range from 45% to 75% and matters most in systems with important, frequently buried residues. Alternative approaches to the solvent access cutoff are also possible. We discuss this fully in the future development of Titr-DMD section and within our test system simulations.

The probability of a protonation state change is assessed for each titratable residue based on instantaneous protonation \( pK_a \) and the residue network information. In this implementation Propka3.1 is used for \( pK_a \) prediction, based on the latest structure from the preceding DMD trajectory. The protonation state change probability is then assessed for each residue. It is calculated differently depending on whether the residue is in a solvent accessible or inaccessible network. For a solvent accessible network the probability is based on the \( pH \) of the solution with which the residue can freely exchange protons (solvent is treated implicitly in DMD). This probability is based off the Henderson-Hasselbalch equation

\[
pH = pK_a + \log \left( \frac{[D]}{[P]} \right)
\]

where \([D]\) is the concentration of the deprotonated state and \([P]\) is the concentration of the protonated state. From this, the probability of adopting the protonated state, \( P_p \), can be defined as

\[
P_p = \frac{[P]}{[P]+[D]} = 10^{\frac{pK_a-pH}{1+10^{pK_a-pH}}}
\]

In the solvent inaccessible case, only the titratable protons already present in the network can be exchanged. The probabilities of protonation state changes for the residues in this contact network are thus coupled; protonation state changes must be determined for the whole network at once, rather than residue by residue. This requires full enumeration of all proton configurations across the network. The preference of a proton to localize on any individual residue is determined by its \( pK_a \), but with comparison to the competing residues in the network rather than the solution \( pH \). To calculate this, let \( R \) be the set of all residues in a network and \( n \) be the number of titratable protons in that network. Let \( T[R] \) be the set of all possible proton configurations \( S, Q, \ldots \) such that \( T[R] = \{ S \in \mathcal{P}(R) : |S| = n \} \). Then for every \( S \in T[R] \) the probability of adopting that proton configuration is

\[
P_C(R,S) = \prod_{q \in T[R] \setminus q} 10^{pK_a,q}
\]

The weighting term for each proton configuration is the product of 10 raised to the \( pK_a \) of each residue that holds a proton in that state \( (s \in S, q \subseteq Q, \ldots) \). This equation is used to calculate the probability of each possible configuration.

Finally, protonation state changes are decided discretely by a Monte-Carlo approach based on the probabilities generated for each network. As with the probabilities, this differs slightly between solvent exposed and buried networks. For solvent accessible networks, a decimal between 0 and 1 is randomly generated for each residue and compared to its decimal probability. If it is above that probability the residue is unprotonated, and if below it is protonated. This approach holds regardless of what the previous protonation state was. For solvent inaccessible networks, the decimal probabilities of all potential protonation configurations are put in a sequential order. A probability range for each configuration, \( S \), is then defined as from \( P_L \) up to \( P_S(R,S) \), where \( P_S \) is the sum of all configuration probabilities already considered and \( P_L(R,S) \) is that of the current configuration. A decimal between 0 and 1 is then randomly generated, and the configuration is decided based on which range this number falls within. Any changes from the previous structure are then made, with hydrogen removed when necessary and DMD placing any new hydrogen on the appropriate heteroatoms. The structure is then ready for the next DMD simulation.

As the Titr-feature may add and remove hydrogen by exchange with implicit solvent, the energies of the structures before and after protonation state changes may not be directly comparable. Therefore, a correction to the DMD energy may be needed at the start of each DMD phase of the simulation. One approach would be to use a value for the solvation energy of a proton, but that can not be obtained directly from experiment and can only be determined by extrapolation.\(^{27,34}\) Values that can be obtained for this (\(-264.3\) and \(-265.9\) kcal/mol\(^{33,46}\)) are large compared to the DMD energy changes associated with structural fluctuations (ca. 100 kcal/mol). Unmodified use of this proton solvation free energy would result in unphysical behavior – Titr-DMD would always deprotonate any residue. Appropriate scaling of the solvation energy is one solution. For this implementation of Titr-DMD an energy correction for each iteration is obtained instead based on the Propka \( pK_a \) of all residues with protonation states that deviate from the original structure. For each protonation state, take the following acid dissociation reaction

\[
PRTN \rightleftharpoons PRTN' + H^+
\]

where \( PRTN \) is the original protein and \( PRTN' \) is the new state. The free energy of this reaction can be written as

\[
\Delta G_{\text{deprot}} = G(PRTN') + G(H^+) - G(PRTN)
\]

Additionally, the \( K_a \) of this reaction is defined as

\[
K_a = 10^{-\Delta G_{\text{deprot}}/RT}
\]

Hence
\[ G(PRTN) + G(H^+) = G(PRTN^-) - RT\ln(10^{-pK_a}) \]

where \( G(PRTN) \) is the uncorrected DMD energy and the left-hand side of the reaction is a corrected energy for a comparable system with the same chemical composition. For the protonation reaction, casting PRTN as \( PRTN^- \) and \( PRTN \) as \( PRTN \) in the original reaction gives the equation

\[ G(PRTN^-) - G(H^+) = G(PRTN^-) + RT\ln(10^{-pK_a}) \]

The energy associated with each protonation state change from the reference structure can therefore be written as \( \pm RT\ln(10^{-pK_a}) \), positive for protonation and negative for deprotonation. These are on the scale of 5-50 kcal/mol, consistent with DMD energy fluctuations. These terms are calculated for each iteration and summed with its DMD energy for the corrected energy.

**Future development of Titr-DMD.** The modularity of Titr-DMD allows for easy adaptation and refinement. Changes to the method do not require reparameterization of the forcefield. Future developments of Propka or any other tool to calculate the instantaneous \( pK_a \) of a protein conformer can be exchanged to generate the probabilities of protonation state change and improve the quantitative accuracy of the feature. The Titr-feature could even be paired with another molecular mechanics method besides DMD, so long as it is in implicit solvent to retain simulation speed. Alternatives to the system-dependent solvent access cutoff are also of interest. One is to use the Propka buried percentage as a scalar probability of solvent accessibility rather than assign a sharp cutoff, while another is based on the solvent-excluded surface (SES) determined by reduced surface. This method defines the contour of the protein that can not be accessed by solvent by rolling a sphere with the van der Waals radius of the solvent (the “probe”) across the protein, avoiding the van der Waals radii of the other atoms. The solvent accessibility of any residue can be determined by measuring the distance of its titratable group to the nearest vertex of the water SES. If the vertex is within the van der Waals radius of the titratable group, it is solvent accessible. We are currently investigating these approaches for future developments.

**Benchmark Systems.** The systems considered for \( pK_a \) prediction have been studied extensively both experimentally and with computational methods (Figure 2A-C). Hen egg-white lysozyme (HEWL) was used as it is a prototypical system for CpHMD benchmarking. The input structure of the protein was taken from the Protein Data Bank (PDB ID 1LZN). All solvent molecules were removed for the simulation — water, nitrogen trioxide, and the sodium ion. As HEWL only reports experimental \( pK_a \) for GLU, ASP, LYS, TYR, and a single HIS residue, both human thioredoxin (HTRX) and human muscle creatine kinase (HMCK) were simulated as well. HTRX brought in CYS and another HIS residue to the dataset alongside many more GLU and ASP. HMCK only added one CYS residue to the dataset, but was included as it is one of the largest proteins with an experimentally identified amino acid \( pK_a \) at 381 residues (compared to 105 residues for HTRX and 129 for HEWL). The initial structure used for HTRX was PDB ID 1ERT, with all water molecules removed and the rotamers labeled ‘A’ used when more than one was recorded. As it is unclear whether the 320-331 loop of HMCK is unstructured or an alpha helix, two structures were used. The unstructured case was based on the A chain of PDB ID 1U6R, mutated back to the WT sequence with the substrate ADP, inhibitor (diaminomethyl-methyl-amino)-acetic acid, all water, nitrogen trioxide, and magnesium ions removed. The alpha loop structure was the same except the 320-331 loop was replaced with the 321-332 loop of the A-chain from PDB ID 3B6R. All experimental reference \( pK_a \) were drawn from the PKAD database. The \( pK_a \) values used ultimately come from Bartik et al. and Webb et al. for HEWL, from Forman-Kay et al. and Qin et al. for HTRX, and Wang et al. for HMCK. The \( pK_a \) predictions from our simulations measure error and deviations to the average of these datasets for each residue with more than one reported value.

The system used to assess pH-conformational coupling was Staphylococcal nuclease mutant V66K, a well characterized system (Figure 2D). Experimental information about protein conformational dynamics, including in the context of \( pH \) change, is difficult to obtain. As discussed in the introduction, the study of SNase mutants is a rare example of this being done. A combination of NMR, CD, and titration suggests that the protonation of LYS66 is concurrent with and may be coupled to the unraveling of the first loop of the alpha helix on which it is located. The V66K mutant was selected as it demonstrates an extreme \( pK_a \) shift of 10.5 down to 5.7 – which alongside the conformational coupling is a real challenge for any CpHMD method. The structure used was PDB ID 2SNM with thymidine-3’,5’-diphosphate, water molecules, and the calcium ion removed.

**Titr-DMD settings.** Benchmarking simulations differ slightly between those done to estimate \( pK_a \) values and those that assess pH-conformational coupling. The \( pH \)-conformational coupling simulations were longer and hotter to achieve the necessary sampling. DMD simulations without the Titr-feature were also run for the \( pH \)-conformation coupling system as a control – to make sure conformational changes are \( pH \) dependent. The \( pK_a \) estimating simulations were run for 2,000,000 DMD timesteps (roughly 100 ns) at 50 K (note that temperature in DMD is defined specifically, and does not di-
The comparison to the solvent access cutoff was done for a 20-Å van der Waals radius of three vertices (one face) of the SES defined as solvent accessible if its amine group is within its on a comparison to the reduced surface SES. LYS66 was then 55%, and 65%) and determined the most physical cutoff based series of different solvent access cutoff (25%, 35%, 45%, and 45% ultimately selected and a temperature of 150K for increased mobility. The other settings were the same as for the predication. The DMD control simulations without the Titr-feature were run for the same time and temperature as the mational coupling simulations.

The exact value of the solvent access cutoff is system dependent and requires special attention/calibration. A cutoff of 75% was originally selected for all of the simulations conducted here. A Propka calculation on the initial structure for each system yielded buried values above ~75% for only the entirely buried residues and those nested in internal folds, while the external surface residues were all well below this percentage. Over the course of the Titr-DMD simulations, only 2 of the 40 residues with experimentally available $pK_a$ in the benchmark set stayed consistently buried (ASP26 in HTRX and CYS282 in HMCK), so these simulations were ultimately not too sensitive to the choice of solvent buried cutoff. This choice was, however, very critical for simulations of $pH$-conformational coupling, as they involved a study of whether the protonation state of a deeply buried residue (LYS66 in a staphylococcal nuclease mutant V66K) is coupled to the dynamics of the protein. In this case, we ran full simulations at a series of different solvent access cutoff (25%, 35%, 45%, 55%, and 65%) and determined the most physical cutoff based on a comparison to the reduced surface SES. LYS66 was then defined as solvent accessible if its amine group is within its van der Waals radius of three vertices (one face) of the SES. The comparison to the solvent access cutoff was done for a 20 ns test Titr-DMD simulation of SNase V66K with a solvent access cutoff of 75%. The solvent accessibility was then calculated for each iteration with a SES generated by Chimera (v 1.13.1).36 The iterations were binned based on their Propka predicted percent buried, and the percent of states deemed to be solvent accessible was calculated for each bin (Figure 3). The results show that for SNase V66K there is a threshold around 65-75% buried where LYS66 is generally buried above that point and not buried below that point. Structures do not, however, become entirely solvent accessible until below 55%. It is important to note that such a low Propka buried % is rare for LYS66 in our short test simulation. We therefore selected for analysis the SNase simulations with solvent access cutoffs of 65% and 45%, one around the inflection point (in Figure 2) and one where LYS66 is only solvent accessible when it is as assured based on our SES test. The results of the other cutoffs are reported in the supporting information (Figure S1, Tables S1-S3).

A total of 45 Titr-DMD and 4 DMD simulations were performed for benchmarking. Simulations were done for HEWL at pH 3, 5, 7, and 9, for HTRX at pH 3, 5, and 7, and for HMCK both with the unstructured and alpha helical 320-331 loop at pH 9. The $pH$ values were selected to straddle the $pK_a$ of residues with experimentally reported values. Simulations were run for SNase at pH 4.6, 5.7, and 7. These values are below, at, and above the experimental $pK_a$ of the LYS66 residue and its coupled dynamic behavior. All are above the denaturing point of the protein. Three replicates were performed for each system and pH. The four DMD simulations were run for SNase to provide a point of comparison. Two were run with LYS66 permanently deprotonated and two with it permanently protonated.

Convergence of the Titr-DMD simulations was attained according to a pair of metrics. This is comprised of the backbone RMSD and the corrected Titr-DMD energy (Figure 4). The RMSD was calculated with the initial structure as the reference and with respect to the alpha carbon and amide nitrogen, carbon, and oxygen of each amino acid. All trajectories come to oscillate around fixed values, indicating convergence.

Figure 3: Comparison of the Propka buried percentage to solvent accessibility according the reduced surface SES. The dark blue line is the fraction of states that are solvent inaccessible according to the SES, where 1 is totally inaccessible and 0 is entirely accessible (left axis). The light blue histogram is the number of states in that bin (right axis). States are binned in units of 5 buried percentage points. Note that most states below 65-70% buried are solvent inaccessible according to the SES and that those below 55-60% buried are entirely inaccessible according to the SES.
Results and Discussion

Titr-DMD offers rapid sampling on limited resources. The combination of DMD and Propka in an implicit solvent makes it a fast and affordable method. We assessed the scaling of Titr-DMD through 1000 step (5 protonation assessments, about 50 ps) simulations of HEWL, HTRX, and HMCK executed with 1, 2, 4, 8, and 16 processors both with and without the Titr-feature. All simulations were run on the same node sequentially during a single submission to reduce the impact of the variability of other demands on the supercomputing cluster. Simulations were performed on AMD Opteron 2380 (2.5 GHz) cores on Hoffman2 at UCLA IDRE. This process was replicated five times, with the average of these results taken (Figure 5). Titr-DMD scales roughly linearly with the number of residues, and scales favorably out to four processors, with additional resources giving diminished returns. The Titr-feature does modestly increase the computational expense of DMD simulations, with the increase in relative runtime over base DMD growing some with the number of processors used.

This largely derives from the need to initialize many short DMD simulations. However, Titr-DMD still runs quite well on limited resources; the CPU time for the four processor tests scales up to 500-1300 CPU hours (or 3-5.5 CPU hours per residue) to reach a 1 ms simulation.

With a couple exceptions, our Titr-DMD method successfully recaptured the experimental $pK_a$ of the test system residues. We calculated the average RMSE between the predicted and experimental values both by type of amino acid and by protein test system (Tables 1-2). The $pK_a$ can be calculated two ways from Titr-DMD, therefore we calculated two average RMSE for each case. Propka $pK_a$ is simply the average of the Propka predicted values from each timestep. The DMD $pK_a$ for a residue is the natural logarithm of the fraction of timesteps in which the residue is protonated. That fraction is analogous to the $K_a$: the relative concentration of the protonated form of the residue. For solvent exposed residues (those that can freely change protonation state just based on their instantaneous $pK_a$) these two $pK_a$ predictions should converge to the same values with appropriate Monte-Carlo sampling. The
results show that indeed the average RMSE are in good agreement between the two methods for each system, with generally small values around or below 1. The only exceptions are for aspartate and cysteine residues. In the case of aspartate it appears that Propka does poorly specifically with residues that have especially depressed values below 2. Cysteine, on the other hand, is handled uniformly poorly by Propka. This may have to do with the fact that both residues are partially buried and one participates in a disulfide bond (which could complicate the Propka prediction and is handled in an ad-hoc manner by DMD). Furthermore, our test set only includes two cysteine residues as experimentally obtained values for this amino acid are rare; the \( pK_a \) predictions for these particular residues could simply be anomalously poor. A different instantaneous \( pK_a \) prediction method than Propka may improve accuracy for these specific residues when necessary for a simulation.

Table 1: Titr-DMD \( pK_a \) RMSE by amino acid compared to experimental values. Each is calculated relative to an aggregate dataset compiled from all considered experimental studies. HEWL, HTRX, and HMCK are all included.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>DMD RMSE</th>
<th>Propka RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASP</td>
<td>1.49</td>
<td>1.44</td>
</tr>
<tr>
<td>GLU</td>
<td>0.47</td>
<td>0.63</td>
</tr>
<tr>
<td>HIS</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>LYS</td>
<td>0.76</td>
<td>0.81</td>
</tr>
<tr>
<td>TYR</td>
<td>0.67</td>
<td>0.52</td>
</tr>
<tr>
<td>CYS</td>
<td>2.28</td>
<td>3.87</td>
</tr>
</tbody>
</table>

Titr-DMD predictions of \( pK_a \) values are competitive with more expensive CpHMD methods. We compare our method on a truncated set of residues that was also assessed by the explicit solvent Vila-Viçosa et al.\(^{28}\) and the Goh et al.\(^{23}\) replica exchange CpHMD methods, and the implicit solvent implementation of the Wallace et al.\(^{27}\) replica exchange CpHMD method. This set includes mostly asparatate and glutamate residues as well as one histidine residue. For this set the Wallace method obtains an RMSE of 0.89 while the Goh method obtains 0.83 and the Vila-Viçosa method obtains 0.83. Our RMSE of 1.47 or 1.45 (for the DMD and Propka \( pK_a \) respectively) is reasonable, as the truncated set of HEWL residues heavily features low-\( pK_a \) aspartate residues, with which our method does worse, and entirely excludes lysine and tyrosine. The RMSE for other residues besides aspartate and cysteine are on par with the more expensive methods.

Table 2: Titr-DMD \( pK_a \) RMSE by protein system compared to experimental values. HEWL* consists of the ASP and GLU residues that are considered in other CpHMD method benchmarks and is comparable to them, while HEWL includes LYS, TYR, and HIS residues as well.

<table>
<thead>
<tr>
<th>System</th>
<th>DMD RMSE</th>
<th>Propka RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEWL</td>
<td>1.30</td>
<td>1.27</td>
</tr>
<tr>
<td>HEWL*</td>
<td>1.47</td>
<td>1.45</td>
</tr>
<tr>
<td>HTRX</td>
<td>1.07</td>
<td>0.93</td>
</tr>
<tr>
<td>HMCK</td>
<td>5.66</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Computational resource scaling benchmark of Titr-DMD, plotted by (A) the number of processors and (B) the number of residues. Note the linear scaling with number of residues and that good performance is reached with four processors. (C) The percent increase of time for Titr-DMD over unmodified DMD. Note that the increase is relatively small and only becomes significant with many processors as the time DMD takes shortens.
Protonation and deprotonation of LYS66 is coupled with loop unraveling according to Titr-DMD. At pH 5.7, the percentage of unraveled states is significantly higher around LYS66 protonation state changes than the total simulation average (Table 4). Moreover, few events at pH 5.7 occur without contemporaneous unraveling. The coupling we observe in our simulations is thus consistent with the experimental hypothesis.32

Table 4: Frequency of unraveling of the SNase mutant 65-69 loop around LYS66 protonation and deprotonation events. ‘Near event’ refers to the percentage of structures within 25 timesteps (before and after) of an event that are unraveled. This value is roughly on par with the total simulation average except at pH 5.7, particularly during the simulation with a 45% solvent access cutoff. ‘By event’ refers to the percent of events that have at least one unraveled structure within 25 timesteps. Again, note that the pH 5.7 simulations show high coupling where protonation state changes nearly always occur alongside some contemporaneous unraveling.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>pH 4.6</th>
<th>pH 5.7</th>
<th>pH 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>45% cutoff</td>
<td>1.11%</td>
<td>30.43%</td>
<td>0.54%</td>
</tr>
<tr>
<td>65% cutoff</td>
<td>3.20%</td>
<td>9.45%</td>
<td>10.97%</td>
</tr>
</tbody>
</table>

Titr-DMD dynamics can predict the pH at which loop unraveling occurs. The Propka predicted pKₐ value of LYS66 is uniformly higher than the experimental 5.7, at an average of 7-8, but still shows a qualitatively correct large drop from the solution value of 10.5 (Table 5). However, the DMD pKₐ is gen-

Figure 6: (A) Criterion for an unraveled 65-69 loop structure in our SNase simulations. This compares the RMSD of the loop (RMSDₜ) to the RMSD of the full protein (RMSDₑ) and compares the distances of important hydrogen bonding contacts (R₁, R₂) to standard values (RHB) to determine structures where the conformation of the loop varies significantly from the original structure. We give the values of the other variables in the main text. (B) Example of a SNase conformation with an unraveled 65-69 loop by our criterion (light blue) overlaid on a structure where it is not unraveled (tan). LYS66 is colored yellow here.

Titr-DMD holds promise for the study of the effect of solution pH on protein structure. Simulations of SNase V66K are qualitatively consistent with rare, experimentally studied dynamics. With the Titr-feature, we observe partial unraveling of the first turn of the alpha helix on which K66 is localized on (residues 65-69), which is not apparent in DMD without titration (Table 3). Unraveling is only observed in 0.002-0.015% of structures in base DMD, while Titr-DMD simulations show it occurs in 3-8% of structures. We define an unraveled state as one where the ALA69-lys66 and ASN68-MET65 hydrogen bonds are broken or breaking and the backbone RMSD of the loop is large relative to that of the backbone, indicative of significant, localized structural change (Figure 6). The criterion is

\[ \frac{S_\text{RMSD}_{\text{loop}}}{(R_1 - R_{\text{HB}}) + (R_2 - R_{\text{HB}})} \leq 2 \]

where RMSD_loop is the total protein and RMSDₑ is for the loop (residues 65-69), R₁ and R₂ are the backbone amide H to carbonyl O distances in A of ALA69-lys66 and ASN68-MET65 respectively, S_RMSD is 2, RHB is 2.5 A (for a long hydrogen bond length), and S_RMSD is 2 A. We only consider structures where the ALA69-lys66 and ASN68-MET65 backbone hydrogen bond distances are both at least 3 A. While unraveling according to our criterion occurs in 3-8% of all states at the appropriate pH, it is not typically sustained for longer than about 1 ns at any one time. We surmise that our simulations do not have enough sampling to capture sustained loop unraveling, but do show the rare events that could lead to it.
erally lower, representing the frequent solvent inaccessibility of the residue. With the solvent access cutoff tuned to 45% it even comes close to the experimental finding. At this cutoff, unraveling is most common in the pH 5.7 simulations and nearly all protonation state changes occur alongside some unraveling. Titr-DMD can qualitatively model coupling between pH and protein structure, and when well calibrated can do so with more quantitative accuracy.

Table 5: Titr-DMD predicted pK\textsubscript{a} for LYS66 in the SNase mutant. These values are averages across all replicate simulations. All of these are above the experimental value, though still significantly shifted down from the solution value of 10.5, except the DMD pK\textsubscript{a} value for the 45% solvent access cutoff simulation. This reflects the typically solvent shielded environment of LYS66 modeled through DMD dynamics that Propka underestimates on its own.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>DMD</th>
<th>Propka</th>
</tr>
</thead>
<tbody>
<tr>
<td>45% cutoff</td>
<td>5.79</td>
<td>7.80</td>
</tr>
<tr>
<td>65% cutoff</td>
<td>7.04</td>
<td>7.96</td>
</tr>
</tbody>
</table>

Conclusions
In this paper we demonstrate Titr-DMD as an effective new method to study pH coupled protein dynamics. The challenges that face any CpHMD method are appropriate conformational and protonation state sampling, accuracy of protonation state changes, and whether the generated conformational ensemble is physically meaningful. Titr-DMD offers great sampling on just a few processors through atomic collision event calculations, implicit solvation, and semi-empirical pK\textsubscript{a} prediction with Propka. Our method obtains reasonably accurate pK\textsubscript{a} predictions for its computational expense. Titr-DMD was successfully benchmarked on the partial unraveling of SNase mutant V66K: one of the few experimentally studied pH coupled conformational changes. Titr-DMD generates a conformational ensemble consistent with experiment, and this ensemble even reflects the experimental pH value of the conformational change. Our method is also modular to further improve sampling and accurate assignment of protonation states. Titr-DMD stands as a promising method to address questions of pH dynamics in industrial catalysis and medicine.

ASSOCIATED CONTENT
Supporting Information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author
*E-mail: ana@chem.ucla.edu. Phone: +1 310 8253769

Funding Sources
Financial support comes from the NSF CHE-1903808 grant to A.N.A., and from the Passan Foundation and US National Institutes of Health grant R35 GM134864 to N.V.D. We also acknowledge UCLA-IDRE and XSEDE for providing computational resources.

Notes
We declare no competing financial interest.

ACKNOWLEDGMENT
We thank the Institute for Digital Research and Education at UCLA and the Extreme Science and Engineering Discovery Environment for supercomputer time. We thank Matthew Hennefarth for advice on integration of the Titr-feature with DMD and a new implementation of DMD. We thank Prof. Feng Ding for advice on the heat exchange parameter of DMD.

ABBREVIATIONS
SNase, staphylococcal nuclease; HTRX, human thioredoxin; HEWL, hen egg-white lysozyme; HMCK, human muscle creatine kinase; SES; solvent-excluded surface.
REFERENCES


