Development & Application of Constant pH Molecular Dynamics (CPHMD$^{MS\lambda D}$) for Investigating pH-mediated Transient Conformational States and Their Effects on Nucleic Acid & Protein Activity

by

Garrett B. Goh

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in the University of Michigan
2015

Doctoral Committee:

Professor Charles L. Brooks III, Chair
Professor Heather A. Carlson
Professor Nils G. Walter
Professor David Sept

Acknowledgements

The author would like to acknowledge the following collaborators and colleagues for their respective contributions (in chronological order):

- Jennifer Knight for the development of multi-site λ-dynamics (MSλD) algorithm and collaboration in the early development of the explicit solvent CPHMD$^{MS\lambda D}$ framework.
- Hashim Al-Hashimi and Evgenia Nikolova for introducing the paradigm of transient conformational states in biomolecular systems, and collaboration on the investigation of Hoogsteen bases in DNA duplexes.
- Ben Hulbert and Jane Zhou for their contribution in the development explicit solvent CPHMD$^{MS\lambda D}$ framework for proteins.
- Elena Laricheva and Afra Panahi for their collaboration and investigation into pH-dependent activity of proteins using the explicit solvent CPHMD$^{MS\lambda D}$ framework.
- Alex Dickson for his contribution in the WExplore sampling method for identification of transiently populated conformational states.
- Nils Walter and Kamali Sripathi for their contribution to AMBER parameterization for nucleobases, and collaboration on the investigation into the catalytic mechanism of the hairpin ribozyme.

Table of Contents

List of Figures

List of Tables

Abstract

pH is a ubiquitous regulator of biological activity, with widespread impact ranging from its role in catalysis to carcinogenesis. Traditionally, a combination of biophysical and computational methods are used to measure pH-dependent activity profiles, and protonation equilibria (i.e. pK_a values) of specific residues, and these data are used in conjunction with structural data to provide mechanistic understanding of pH-mediated biological function. More recent developments have also demonstrated the role of transient conformational states in a wide range of biological activities, which naturally leads to the question of how pH affects such transient states, and in turn, their resulting functional activity. In the study of biomolecular transient states, the detection limit is the key limitation of most experimental techniques. To bridge the gap in detection limit, we have developed an appropriate molecular dynamics based computational method, where protonation states are dynamically adjusted as a function of an external pH bath and the local environment surrounding the titrating site. Also known as the explicit solvent constant pH molecular dynamics ($\text{CPHMD}^{\text{MS}\lambda\text{D}}$) framework, we use $\text{CPHMD}^{\text{MS}\lambda\text{D}}$ simulations and enhanced sampling methods to demonstrate the role of pH-regulated transient states in both nucleic acid and protein activity. First, we demonstrate the utility of $\text{CPHMD}^{\text{MS}\lambda\text{D}}$ simulations in conjunction with NMR experiments to characterize transiently populated Hoogsteen GC^+ base pairs in DNA duplexes. The role of pH-dependent transient states is then generalized to RNA activity, including that of the catalytic mechanism of the hairpin ribozyme, where the existence of pH-dependent transient states can be used to reconcile a collection of seemingly inconsistent experimental observations in the literature. In addition, our $\text{CPHMD}^{\text{MS}\lambda\text{D}}$ simulations of proteins have elucidated the role of pH-dependent transient states in residues that are buried or occluded from solvent, including that of the pH-dependent optical properties of a cyan fluorescent protein mutant, where the existence of pH-dependent transient states can be used to explain its non-monotonic spectroscopic behavior.

Chapter 1: Introduction

1.1 The role and importance of pH

pH is one of the critical regulators of biological activity. Enzymatic activity is optimized within a narrow pH range,[1] often requiring the participation or presence of ionizable residues such as aspartic acid, glutamic acid and/or histidine in the active site,[2] and accurate measurement of their pK_a values is crucial in understanding the catalytic mechanism.[3-5] In recent years, the role of pH regulation in nucleic acid systems has been acknowledged,[6,7] where parallels to proteins can be drawn, such as the catalytic activity of ribozymes (ribonucleic acid enzymes),[8-13] demonstrating the ubiquity of pH regulation in biological processes. Apart from its influence on catalytic activity, pH regulation has been observed in numerous other processes including protein folding,[14-17] protein-protein interactions,[18] protein-substrate binding,[19,20] translational recoding,[21] and aberrant pH regulation has even been implicated in cancer-related physiology.[22] As such, specific examples of pH-dependent properties encompass a wide variety of systems, such as the catalytic mechanism of dihydrofolate reductase,[23] proton gradient driven ATP synthesis,[24] and the influenza virus haemagglutinin.[25]

Histidine, with a pK_a of 6.5, is perhaps the most commonly implicated residue-of-interest in pH-dependent processes. While the pK_a values of amino acid monomers have been known for decades, the microenvironment around the residue located in a protein environment may alter its pK_a value, shifting them towards physiologically relevant pH conditions. Similarly, in nucleic acids, questions about the role of protonated nucleotides with shifted pK_a values in modulating the pH-dependent properties of nucleic acids have emerged.[6,7] Protonated nucleotides are now known to serve as key catalytic residues in many ribozymes,[8-13] and are implicated in the pH-dependent biological activities of numerous RNA systems, such as the retrovirus pseudoknot,[26] peptidyl-transferance center of the ribosome,[27-32] helix 69 of the 50S ribosomal subunit,[33,34] and the U6 intramolecular stem-loop of the spliceosome complex.[35] In DNA, the presence of protonated A^+•C base pairs are known to cause mutagenic and carcinogenic effects.[36,37]

Measuring the pH-dependent activity profile has allowed investigators to determine the overall, or macroscopic, pK_a of the biological process, but limited information can be gleamed about the specific residues that control such pH-dependent activity. Thus, the ability to measure the site-specific pK_a of a particular residue is invaluable in identifying key titrating residues and understanding the mechanism of these pH-dependent biological processes. To achieve this goal, biophysical experimental techniques used to study pH-dependent behavior in both proteins and nucleic acids have made significant progress over the last decade, notably in the use of NMR spectroscopy to measure site-specific pK_a values.[38,39] More recent approaches also include the use of nucleobase analogs that serve as pH-dependent fluorescent sensors,[40-42] and the use of Raman spectroscopy on a crystallized enzyme.[43,44] Despite the copious amounts of biochemical and structural data emerging from these studies, in some systems there still remains ambiguity as to the specific function of these protonated residues.

1.2 Transiently Populated Conformational States in Structural Biology

Until recently, the dominant approach to probe the mechanism of pH-dependent activity has focused on the analysis of static structures typically solved near physiological pH. Recent studies have demonstrated the increasingly important role of biomolecular transient conformational states, hereon referred to as “transient states”, in a wide range of biological activity, from protein folding[45] to ligand binding.[46] Such transient states are typically minor populations that comprise between 0.1% to a few percent of the total population under physiological conditions,[47] but are nevertheless critical in providing a more comprehensive understanding of the processes that they govern.

In the context of our investigation, this naturally leads to the question of how pH affects such transient states, and in turn, their resulting activity. While the list of examples of pH-dependent transient states is still growing, there is precedence of their importance, such as in membrane fusion involving the influenza hemagglutinin HA2 subunit as suggested by Bax and co-workers,[48] although detailed experimental characterization of these pH-dependent transient states has yet to be reported owing to the fact that they border the detection limits of experimental methods. In addition, Al-Hashimi and co-workers have also recently discovered the existence of low population transient state conformations that are functionally important in both RNA and DNA systems, some of which are known to exhibit pH-dependent behavior,[49,50] demonstraing the ubiquity of transient states across different classes of biomolecules. Conventional experimental techniques such as NMR spectroscopy,[38,39] pH-dependent fluorescent nucleobase analogs[40-42] and Raman spectroscopy[43,44] have not been able to directly characterize such pH-dependent transient states, although progress has been made through the development of novel relaxation dispersion NMR spectroscopy techniques,[51,52] and room temperature X-ray crystallography.[53] has made this endeavor more plausible. Probing pH-dependent transient states will undoubtedly prove to be more challenging, as not only are we dealing with low population states, but it is also necessary to deconvolute the pH effects.

1.3 Existing Computational Tools to Model pH Effects

Using computational tools to augment and inform the experimental investigation of pH-dependent processes, notably in dealing with systems that include transient states that are at the forefront of experimental detection limits can be advantageous. Traditionally, molecular dynamics (MD) simulations, which have the ability to provide detailed atomistic insight from first principles, can be used to model the effects of high and low pH on the resulting biomolecule, and in some cases have been used to shed light on existing ambiguities in pH-mediated activity.[54-58] However, conventional MD simulations are only capable of modeling fixed protonation states, and are limited by the fact that substantial prior knowledge about the identity of the key residues and their corresponding protonation states is required. As simulations would be most useful in situations where there is a lack of experimental data, it is evident that the impact of traditional MD simulations alone will be rather limited.

The absence of experimental pK_a values may be resolved by computing pK_a values by using a better theoretical treatment of the electrostatics in proteins and nucleic acids. One of the more successful approaches is that based on the Poisson-Boltzmann (PB) equation methodology, which has achieved reasonable success in predicting protein pK_a values.[59] In terms of predicting the pK_a values of nucleic acids, Honig, Pyle, and co-workers have also recently demonstrated its feasibility using the non-linear Poisson–Boltzmann (NLPB) equation.[60] A key limitation of initial PB methods was the lack of conformational flexibility, although this has been partially addressed using approaches like tuning the effective protein dielectric constant[61] and including representations of multiple conformations.[62,63] The need for conformational flexibility led to the development of the other major physics-based approach in computational pK_a predictions, which is based on traditional molecular dynamics (MD) simulation. Warshel and co-workers were the first to demonstrate the use of free energy calculations to calculate the pK_a values of protein residues.[64-67] Subsequent developments in the MD community have sought to couple the protonation state of the titrating residue with the dynamics of the protein itself. Such pH-coupled simulations, which have been termed constant pH molecular dynamics (CPHMD), are uniquely suited to model realistic pH-dependent responses, even in systems where there is limited experimental data because no *a priori* information on the identity of key titrating residues and their protonation state is required, making them uniquely suited to investigate pH-dependent transient states and other systems where there is limited experimental data. In this formalism, the protonation states of titrating residues change dynamically throughout the simulation that is set according to the external pH bath, and further adjusted according to the changes in the electrostatic microenvironment around the titrating residue. Unlike the pK_a values calculated from traditional computational methods like those based on the PB equation, the CPHMD framework has the added advantage of including dynamical information to its free energy calculations, making it more suitable for modeling pH-dependent properties that correlate to structural fluctuations or local conformation changes.

The CPHMD methodology has been implemented using two distinct approaches, which vary in the manner in which the titration coordinates are treated – either discretely or continuously.[68] In the discrete CPHMD variant, the MD sampling of atomic coordinates is combined with the Monte Carlo (MC) sampling of protonation states. At regular intervals during a typical MD simulation, a MC step is performed to determine the change of the protonation state. Discrete CPHMD was first reported by Bürgi *et. al*.,[69] which was computationally expensive at that time and suffered from convergence issues, owing to the fact that it was performed in explicit solvent and used the more expensive thermodynamic integration approach to calculate the energies used in the MC evaluation step. Baptista and co-workers reported a similar discrete CPHMD implementation but used the Poisson-Boltzmann finite-difference method to calculate the energies used in the MC evaluation step.[70-72] With the advances in implicit solvation models around this time,[73,74] and the initial convergence issues reported for explicit solvent CPHMD,[69] subsequent developments in discrete CPHMD by Dlugosz and Antosiewicz,[75,76] and Mongan *et. al.*,[77] were implemented using a Generalized-Born (GB) implicit solvent model. More recent improvements in the discrete CPHMD community have been focused on achieving better sampling by enhanced sampling techniques, such as Accelerated Molecular Dynamics by Williams *et. al*.[78] and replica exchange strategies by Roitberg and co-workers.[79-81] Others in the field, namely Warshel and co-workers, have focused on developing a more physically realistic form of CPHMD, using time-dependent MC sampling of the proton transfer process,[82] and the empirical valence bond (EVB) framework to simulate proton transfer between solute and solvent.[83]

By contrast, in the continuous CPHMD variant, which was first reported by Baptista *et. al.*[84] and Borjesson *et. al.*,[85] titration coordinates can be treated as mixed states. In the continuous CPHMD variant developed by Brooks and co-workers, the titration coordinate represents an instantaneous microstate, and it is propagated continuously between the protonated and unprotonated states using the λ dynamics approach.[86-88] Continuous CPHMD allows one to avoid sudden jumps in potential energy that occur after a successful MC move in the discrete CPHMD variant, and potentially avoids artifacts that may be caused by the MC moves in titration coordinates. Additionally, continuous CPHMD facilitates coupled proton moves, which would need to be engineered as specific move types in the MC-based variant. Continuous CPHMD was originally implemented in implicit solvent,[89] improved to account for proton tautomerism,[90] and it provided the first demonstration of using enhanced sampling strategies to accelerate sampling and convergence in CPHMD simulations.[91] The effectiveness of continuous CPHMD has been demonstrated on numerous pH-dependent systems, including of protein folding,[92,93] aggregation of Alzheimer's beta-amyloid peptides,[94] pH-triggered chaperon activity of HdeA dimers,[95] electrostatic effects on protein stability,[96] self-assembly of spider silk proteins,[97] and RNA silencing in the carnation italian ringspot virus.[98] Other investigators in the field have also seen a number of successes using discrete CPHMD simulations.[99-101]

1.4 Explicit Solvent Constant pH Molecular Dynamics

While the move to implicit solvent CPHMD has obvious advantages in sampling and convergence, a number of unresolved issues have emerged over the years. It has been reported that the generalized Born implicit solvent model underestimates the desolvation of buried charge-charge interactions,[91] causing a systematic overstabilization of the ionized form[102] and consequently increasing the error of predicted pK_a values. In addition, these models are known to cause structural compaction which may distort the overall structure,[96,103] introducing another source of error in modeling pH-dependent dynamics. Furthermore, in systems such as ion channels[104-106] and some transmembrane proteins,[107] where the microscopic interactions of discrete ions and water with the protein are important, the use of an explicit solvent representation of the solvent environment is desirable. Furthermore, for nucleic acids, existing implicit solvent model have only reported success on the most basic A-form RNA or B-form DNA structure structures,[108] and more esoteric structural features, which are typically present in most RNA structures that are implicated to pH-mediated activity may not be modeled correctly.

Therefore, there is an impetus to re-introduce explicit solvent into the CPHMD framework. At the time when this dissertation first started in 2010, the only explicit solvent

CPHMD reported was by Grubmüller and co-workers that used the continuous CPHMD variant, but that work was limited to a proof of concept demonstration for model amino acid compounds, and no practical applications for larger full-sized proteins were reported.[109] At about the same time, Wallace and Shen reported a hybrid solvent continuous CPHMD model, where the Cartesian coordinates of the protein was propagated in explicit solvent, and the titration coordinates were propagated using the GB implicit model, and it was shown to reduce the errors introduced by the implicit CPHMD framework.[103] In a related work by Roitberg and co-workers in subsequent years, the hybrid CPHMD framework was also implemented using the discrete CPHMD variant, and it too has been demonstrated to reduce the errors associated with the implicit solvent model, notably for longer timescale CPHMD simulations.[110]

1.5 Dissertation Outline

In this dissertation, we report on the first viable explicit solvent CPHMD framework based on the newer multi-site λ-dynamics algorithm (MSλD) to model pH-dependent dynamics. **Chapter 2** summarizes the theory and methodology behind the explicit CPHMD$^{MS\lambda D}$ framework. Early CPHMD$^{MS\lambda D}$ work (**Chapter 3.1**) was first tested on simple nucleotide compounds,[111] before proceeding to simulate full-sized RNA systems.[112] Initial sampling challenges were identified, which were alleviated through the use of a pH-replica exchange (pH-REX) enhanced sampling method (**Chapter 3.2**).[113] The explicit CPHMD$^{MS\lambda D}$ framework, with pH-REX sampling improvements was later extended to model pH effects in proteins (**Chapter 3.3**),[114] and adopted to the AMBER force field (**Chapter 3.4**) for use with RNA systems with more complicated topology.[115,116]

In collaboration with NMR studies, CPHMD$^{MS\lambda D}$ simulations were used to characterize transient Hoogsteen GC$^+$ base pairs in DNA duplexes (**Chapter 4.1**).[117] Subsequent work led to the investigation of pH-mediated transient states and their effect RNA activity, notably on the hairpin ribozyme (**Chapter 4.2**),[115,116] where a combination of CPHMD$^{MS\lambda D}$ and WEXplore,[118] a hierarchical weighted ensemble sampling technique, were used to identify pH-dependent transient states, which were critical in reconciling seemingly baffling and/or conflicting experimental observations. In the context of protein systems, the pH-dependent dynamics of buried ionizable groups in staphylococcal nuclease were simulated by CPHMD$^{MS\lambda D}$ simulations, where the role of pH-dependent transient states was first elucidated (**Chapter 5.1**).[119] Subsequent work led to the investigation of the unusual pH-dependent optical properties of a mutant of cyan fluorescent protein (CFP),[120] where the identified pH-dependent transient states were pivotal in explaining its non-monotonic optical properties (**Chapter 5.2**).

Chapter 2: Explicit Solvent Constant pH Molecular Dynamics: Theory & Methods

2.1 Theory

Note: Chapter 2.1 was adapted from the following references.[111-114]

2.1.1 Constant pH Molecular Dynamics Framework

We briefly review the theory behind constant pH molecular dynamics (CPHMD). In CPHMD, the protonation state of the titrating residue is described by a continuous variable, λ. In the original implementation of continuous CPHMD, the dynamics of λ is described according to λ-dynamics, a formalism that couples the dynamics of λ to the dynamics of the protein system. The simulation is under the influence of a hybrid Hamiltonian and its potential energy is described by:

$$U_{tot}(X,\{x\},\{\lambda\}) = U_{env}(X) + \sum_{\alpha=1}^{N_{sites}} \sum_{i=1}^{2} \lambda_{\alpha,i} \left(U\left(X, x_{\alpha,i}\right)\right)$$

$$+ \sum_{\alpha=1}^{M_{sites}-1} \sum_{i=1}^{N_S} \sum_{T=\alpha+1}^{M_{sites}} \sum_{j=1}^{N_T} \lambda_{\alpha,i}\lambda_{T,j} \left(U\left(x_{\alpha,i}, x_{T,j}\right)\right) \qquad (2.1.1.1)$$

where N_{sites} is the total number of titrating residues (α), which has *i* number of protonation states (typically 2). *X* represents the coordinates of the environment atoms (i.e., the parts of the protein, solvent, etc… that are not titrating). Both $x_{\alpha,1}$ and $x_{\alpha,2}$ represent the coordinates of atoms in residue α that are associated with the protonated and unprotonated states, respectively. The titrating proton and the other atoms whose charges vary according to the protonation state of the residue (usually atoms within 2-3 bonds from the titrating proton) are included in both $x_{\alpha,1}$ and $x_{\alpha,2}$ and are defined as a part of the "titrating fragment.". λ serves as a scaling factor that is associated with each titrating residue α and its value describes the physically relevant protonated ($\lambda_{\alpha,1}$ = 1) and unprotonated ($\lambda_{\alpha,2}$ = 1) states. The double summation signifies the interaction between the environment and all protonation states at each site, and the third term designates the interaction of the protonation states at one site with the protonation states at another site. Protonation states at each site are independent and do not interact with each other in the simulation.

In the explicit solvent CPHMD$^{MS\lambda D}$ simulation framework developed, we utilized the improved λ^{Nexp} functional form of λ implemented under multi-site λ-dynamics (MSλD).[121,122] The scaling factor that is associated with the titrating residue α changes dynamically throughout the simulation and is described by a set of continuous coordinates that are governed by the following equations:

$$\lambda_{\alpha,i}^{N\exp} = \frac{e^{c\sin\theta_{\alpha,i}}}{\sum_{j=1}^{N} e^{c\sin\theta_{\alpha,j}}} \qquad (2.1.1.2)$$

When applied to the two-state system representing the protonated and unprotonated forms this functional form becomes:

$$\lambda_{\alpha,1} = \frac{e^{c\sin\theta_{\alpha,1}}}{e^{c\sin\theta_{\alpha,1}} + e^{c\sin\theta_{\alpha,2}}} \quad \text{and} \quad \lambda_{\alpha,2} = \frac{e^{c\sin\theta_{\alpha,2}}}{e^{c\sin\theta_{\alpha,1}} + e^{c\sin\theta_{\alpha,2}}} \qquad (2.1.1.3)$$

This new form implicitly satisfies the constraints as required by λ-dynamics:

$$0 \le \lambda_i \le 1 \quad \text{and} \quad \sum_{i=1}^{N} \lambda_i = 1 \qquad (2.1.1.4)$$

The use of the λ^{Nexp} functional form also expands the future functionality of the explicit solvent CPHMD$^{MS\lambda D}$ framework to titrate between more than two states, such as the tautomeric forms of titrating groups.

2.1.2. Calibrating CPHMD Simulations

CPHMD$^{MS\lambda D}$ simulations are calibrated on model compounds (i.e., amino acids or nucleosides) to reproduce the external pH environment. Modeling of the external pH bath is achieved by introducing a fixed biasing potential parameter ($\mathrm{F}_{\alpha,2}^{\mathrm{fixed}}$) to the unprotonated state, which results in the following biased potential energy function:

$$U_{tot}(X,\{x\},\{\lambda\}) = U_{env}(X) + \sum_{\alpha=1}^{N_{sites}} \sum_{i=1}^{2} \lambda_{\alpha,i}\left(U\left(X, x_{\alpha,i}\right) - F_{\alpha,i}^{fixed}\right)$$
$$+ \sum_{\alpha=1}^{M_{sites}-1} \sum_{i=1}^{N_S} \sum_{T=\alpha+1}^{M_{sites}} \sum_{j=1}^{N_T} \lambda_{\alpha,i}\lambda_{T,j}\left(U\left(x_{\alpha,i}, x_{T,j}\right)\right) \qquad (2.1.2.1)$$

For the initial calibration, the free energy of deprotonation ($\Delta\mathrm{G}_{\mathrm{protonation}}$) of each isolated model compound calculated using traditional λ-dynamics. The free energy of protonation ($\Delta\mathrm{G}_{\mathrm{protonation}}$) is used to calibrate the biasing potential applied to the unprotonated state ($\mathrm{F}_{\alpha,2}^{\mathrm{fixed}}$) that simulates the effect of an external pH environment, and the other fixed biasing potential applied to the protonated state ($\mathrm{F}_{\alpha,1}^{\mathrm{fixed}}$) is kept at zero. By setting the value of $\mathrm{F}_{\alpha,2}^{\mathrm{fixed}}$ to $\Delta\mathrm{G}_{\mathrm{protonation}}$, approximately equal populations of protonated and unprotonated states are sampled in the simulation. Under this condition, the external pH environment is equal to the $\mathrm{pK_a}$ value of the model compound. To change the pH of the simulation, $\mathrm{F}_{\alpha,2}^{\mathrm{fixed}}$ can be adjusted by the following equation:

$$F_{\alpha,2}^{fixed} = \Delta\mathrm{G}_{\mathrm{protonation}} + \ln(10) k_B T (\mathrm{pK_a} - \mathrm{pH}), \qquad (2.1.2.2)$$

where pH is the external pH of the simulation and $\mathrm{pK_a}$ is the experimental $\mathrm{pK_a}$ of the model compound. The fixed biasing potential is pre-calculated and its value, corresponding to the specified external pH, is universally applied to all residues of the same type regardless of the local microenvironment it is in.

In explicit solvent CPHMD$^{MS\lambda D}$ simulations, when the titration coordinates are allowed to propagate dynamically, the two end points that correspond to physical protonation states may not be sufficiently sampled to yield converged estimates of the pK$_a$ shifts. To ameliorate this issue, the inclusion of an extra variable biasing potential (F^{var}) is introduced, which can be adjusted to tune the sampling efficiency of titration coordinates and the fraction of physical protonation states:

$$F_{\alpha,i}^{\mathrm{var}} = \begin{cases} k_{bias}\left(\lambda_{\alpha,i} - 0.8\right)^2; & if\,\lambda_i < 0.8 \\ 0; & otherwise \end{cases} \tag{2.1.2.3}$$

Thus, in the CPHMD treatment, titratable groups in proteins may be viewed as model compounds that are perturbed by the introduction of the local environment.

2.1.3. pH Replica Exchange Enhanced Sampling

The potential for slow convergence of protonation state sampling in CPHMD simulations has been well documented, and is exacerbated for residues with conformationally-coupled pK$_a$ values, where they undergo a local conformation change that causes them to sample different electrostatic environments yielding distinct microscopic pK$_a$ values.[78,112,123] Early work by Brooks and co-workers on protein CPHMD simulations has demonstrated that the introduction of a temperature replica exchange (T-REX) protocol can significantly accelerate sampling to address such issues.[91] However, using T-REX in explicit solvent MD simulations typically incurs a large computational expense, for example, a moderate sized protein of ~100 residues (40k atoms when solvated) requires at least 20 replicas to achieve reasonable exchange rates between adjacent temperature replicas, and when simulating CPHMD across a reasonable pH range (e.g., pH 5 to 9), the total number of replicas required increases to ~100. Therefore, we used a pH replica exchange (pH-REX) sampling strategy instead, and the pH-REX sampling protocol implemented in our work is based on the work of Wallace and Shen,[103] where simulations performed at various pH conditions are exchanged based on the following Metropolis criterion:

$$P = \begin{cases} 1; & \text{if } \Delta \le 0 \\ \exp(-\Delta); & \text{otherwise} \end{cases} \quad \text{where} \quad \Delta = \beta \begin{bmatrix} U^{\mathrm{pH}}(\{\lambda_i\}; \mathrm{pH}') + U^{\mathrm{pH}}(\{\lambda_i'\}; \mathrm{pH}) \\ -U^{\mathrm{pH}}(\{\lambda_i\}; \mathrm{pH}) - U^{\mathrm{pH}}(\{\lambda_i'\}; \mathrm{pH}') \end{bmatrix} \tag{2.1.3.1}$$

where β is 1/k$_b$T, the first two terms, $U^{\mathrm{pH}}(\{\lambda_i\}; \mathrm{pH}')$ and $U^{\mathrm{pH}}(\{\lambda_i'\}; \mathrm{pH})$ are the pH-biasing potential energies for the two adjacent replicas after the exchange, and the next two terms, $U^{\mathrm{pH}}(\{\lambda_i\}; \mathrm{pH})$ and $U^{\mathrm{pH}}(\{\lambda_i'\}; \mathrm{pH}')$ are the corresponding energies for the respective replicas before the exchange.

2.2 Methods

Note: Chapter 2.2 was adapted from the following references.[111-114,116]

2.2.1. Simulation Methods: Structure Preparation

Input structures of the model compounds (i.e. peptides, nucleosides) and test compounds (i.e. dipeptide, dinucleotide sequences) were generated from the CHARMM topology files using the *IC* facility in CHARMM while hydrogen atoms were added using the *HBUILD* facility.[124] Model and test compounds were solvated in a cubic box of explicit TIP3P water[125] using the convpdb.pl tool from the MMTSB toolset.[126] For each system, it was first neutralized, before an appropriate number of Na^+ and Cl^- counterions was added to match the experimental ionic strength. For the mononucleotides, two isomers in the form of 5'-phospate and 3'-phosphate were constructed using the patch keywords *5PHO* and *3PHO* respectively, in CHARMM. All other nucleic acid structures had hydroxyl groups patched to the terminal ends via patch keywords *5TER* and *3TER*. Peptides and protein structures were capped at the N-terminus and C-terminus using CHARMM's *ACE* and *CT2* patches.

Additional patches were constructed to represent the protonated forms of nucleic acids and amino acids. All of the associated bonds, angles and dihedrals were explicitly defined in the patch. The environment atoms were defined as all atoms that were not included in the titratable fragments. Each titratable residue was simulated as a multiple topology model that explicitly included atomic components of both the protonated and unprotonated forms. The CHARMM parameters for the partial charges of aspartic acid, glutamic acid and lysine used in this study were reported previously by Lee *et. al.*[89] Partial charges for the three protonation states of histidine were obtained without modification from the HSP, HSE and HSD residues as reported in the CHARM22 all-atom force field for proteins.[127] Parameters for nucleic acid were derived as indicated in **Chapter 3.1** and **Chapter 3.4**.

Input structures of full-sized biomolecules were prepared in a similar fashion as that used for the model and test compounds. The input structure for the protein hen egg-white lysozyme (HEWL), the 45-residue binding domain of 2-oxoglutarate dehydrogenase multienzyme complex (BBL) and the 56-residue N-terminal domain of ribosomal L9 protein (NTL9) were generated from the PDB file (accession codes: 2LZT, 1W4H, 1CQU respectively). The input structure for lead-dependent ribozyme was generated from the PDB file (Accession code: 1LDZ), using the lowest energy NMR structure reported.[128] The excised GAAA tetraloop was constructed by extracting residues 12 to 21 from the lead-dependent ribozyme, and harmonic distance restraints were applied to enforce base pairing between residues A12 and U21.

2.2.2. Simulation Details: Molecular Dynamics

MD simulations were performed within the CHARMM macromolecular modeling program (version c36a6) using either (i) CHARMM36 all-atom force field for RNA[129], (ii) modified AMBER force field for RNA, (iii) CHARMM22 all-atom force field for proteins[127] or (iv) CHARMM36 all-atom force field for proteins[130,131] and TIP3P water.[125] The simulation set up for λ dynamics is similar to that reported by Knight and Brooks.[121,122] The SHAKE algorithm[132] was used to constrain the hydrogen-heavy atom bond lengths. The Leapfrog Verlet integrator was used with an integration time step of 2 fs. A non-bonded cutoff of 12 Å was used with an electrostatic force shifting function or force switching function (in latter studies) and a van der Waals switching function between 10 Å and 12 Å. The distance cutoff in generating the list of pairwise interactions was 15 Å. While group-based 8 Å cutoffs investigated in the 1990s were notoriously poor in reproducing accurate dynamics of biomolecules relative to the Ewald summation technique,[133,134] modern atom-based cutoff schemes with sufficiently long cutoff distances (12 Å),[135] such as those employed in this study, have been shown to be comparable to the Ewald summation technique in modeling the dynamics of both proteins[136] and nucleic acids.[137] The threshold value for assigning $\lambda_{\alpha,i}$ =1 was $\lambda_{\alpha,i} \geq 0.8$. Variable biases ($F^{var}$) were added to the hybrid potential energy function and the associated force constant (k_{bias}) was optimized to enhance transition rates between the two protonation states. Since identical k_{bias} values were applied to both protonated and unprotonated states, the PMF at the end-points were not altered and no reweighting scheme was required.

CPHMD simulations utilize an extended Hamiltonian approach, where the protonation state of the residue is described by a continuous variable, λ, which is propagated simultaneously with the spatial coordinates at a specified external pH using multi-site λ-dynamics. The $CPHMD^{MS\lambda D}$ simulations performed in the multi-site λ-dynamics framework (MSλD)[121,122] within the BLOCK facility, using the λ^{Nexp} functional form for λ (*FNEX*) with a coefficient of 5.5.[121,122] The titratable fragment included the protonation site and adjacent atoms whose partial charge differed according to the protonation state. The environment atoms were defined as all atoms that were not included in the titratable fragments. Linear scaling by λ was applied to all energy terms except bond, angle and dihedral terms, which were treated at full strength regardless of λ value to retain physically reasonable geometries. Each θ_α was assigned a fictitious mass of 12 amu•$Å^2$ and λ values were saved every 10 steps. The temperature was maintained at 298K by coupling to a Langevin heatbath using a frictional coefficient of $10ps^{-1}$.

After an initial minimization, most systems were heated for 100 ps and equilibrated for 100ps to 400ps. This was followed by a production run of variable length, ranging from 3 ns to >50 ns per system or replica. For most systems, $CPHMD^{MS\lambda D}$ simulations were performed across the pH range, with integer value pH spacing, as indicated in the titration curves. Larger systems with more titratable groups used a smaller spacing of 0.5 pH intervals. In the pH-REX simulations, exchange attempts were made at every 1 to 2 ps. All CPHMD$^{MS\lambda D}$ simulations were performed in triplicate.

2.2.3. Calculating pK$_a$ values

The populations of unprotonated (N^{unprot}) and protonated (N^{prot}) states are defined as the total number of times in the trajectory where conditions $\lambda_{\alpha,1} > 0.8$ and $\lambda_{\alpha,2} > 0.8$ are satisfied respectively. They are used in the calculation of the fraction of physical states, which is the ratio of N^{unprot} and N^{prot} states over all states (which include intermediate λ values). The unprotonated fraction (S^{unprot}) is calculated for each pH window:

$$S^{unprot}(\mathrm{pH}) = \frac{N^{unprot}(\mathrm{pH})}{N^{unprot}(\mathrm{pH}) + N^{prot}(\mathrm{pH})} \quad \text{(u)}$$

S^{unprot} values computed across the entire pH range, were then fitted to a generalized version of the Henderson-Hasselbalch (HH) formula[138] to obtain a single pK$_a$ value:

$$S^{unprot}(\mathrm{pH}) = \frac{1}{1+10^{-n(pH-pKa)}} \quad (2.2.3.2)$$

Unless specified otherwise, the reported pK$_a$ value and its uncertainty correspond to the mean and standard deviation calculated from 3 sets of independent runs. The pK$_a$ values and the Hill coefficients (n) were calculated using **equation 2.2.3.2**. In this formalism, n has a theoretical value of one and deviations from this value indicate the degree of cooperativity (n > 1) or anti-cooperativity (n < 1) between strongly interacting titratable groups.[138,139] In the calculation of transition rates, a transition is defined as a move in λ space between physical protonation states using the same definitions for calculating N^{unprot} and N^{prot} (i.e., moving between $\lambda_{\alpha,1} > 0.8$ and

$\lambda_{\alpha,1} < 0.2$ constitutes a valid transition). The transition rate statistics reported are calculated from the simulation where the pH value was closest to the pK$_a$ value of the residue in question.

2.2.4. Treatment of Symmetrical Systems

In some dipeptide test compounds, the symmetry of the system may render the environment around each titrating residue to be similar. In such a situation of coupled titrating residues, protonation state statistics for a specific residue may not be associated with the titrating residue. Therefore, the pK$_a$ calculation has to be performed using a modified version of **equation 2.2.3.2**, where the combined S$^{\text{unprot}}$ ratio for all *i* residues is fitted to the following equation:

$$\sum_i^N S_i^{unprot}(\text{pH}) = \sum_i^N \frac{1}{1+10^{-(pH-pKa_i)}} \qquad (2.2.4.1)$$

Derivation of Equation 2.2.3.2 from Mean Field Approximation

Here, we show how **equation 2.2.3.2** can be derived from the mean field approximation (i.e. equation 1b) from Bashford and Karplus.[140]

$$\log\frac{\theta}{1-\theta} = pKa - pH \qquad (2.2.4.2)$$

where θ is the probability that the site is protonated:

$$\theta = \frac{N_p}{N_p + N_u} \qquad (2.2.4.3)$$

Therefore **equation 2.2.4.2** can be rewritten as:

$$\log\frac{\theta}{1-\theta} = \log\frac{\left(\dfrac{N_p}{N_p+N_u}\right)}{\left(\dfrac{N_u}{N_p+N_u}\right)} = pKa - pH \qquad (2.2.4.4)$$

$$\log\frac{N_p}{N_p+N_u} = \log\frac{N_u}{N_p+N_u} + pKa - pH \qquad (2.2.4.5)$$

$$\log \frac{N_p}{N_p + N_u} = \log \frac{N_u}{N_p + N_u} + \log\left(10^{pKa-pH}\right) \tag{2.2.4.6}$$

$$\frac{N_p}{N_p + N_u} = \frac{N_u}{N_p + N_u}\left(10^{pKa-pH}\right) \tag{2.2.4.7}$$

$$\frac{N_u}{N_p + N_u} + \frac{N_u}{N_p + N_u}\left(10^{pKa-pH}\right) = 1 \tag{2.2.4.8}$$

$$\frac{N_u}{N_p + N_u}\left(1 + 10^{pKa-pH}\right) = 1 \tag{2.2.4.9}$$

$$\frac{N_u}{N_p + N_u} = \frac{1}{1+10^{pKa-pH}} = \frac{1}{1+10^{-(pH-pKa)}} \tag{2.2.4.10}$$

which is the same form as **equation 2.2.3.2** for n = 1:

$$S = \frac{N_u}{N_p + N_u} = \frac{1}{1+10^{-(pH-pKa)}} \tag{2.2.4.11}$$

Derivation of Equation 2.2.4.1 from Decoupled Site Representation

Here, we show how **equation 2.2.4.1** can be derived from the decoupled site representation (DSR) from Onufriev *et. al.*[138] We start with the following expression:

$$\sum_i^N \langle x_i \rangle = \sum_j^N \frac{10^{pKa_j-pH}}{1+10^{pKa_j-pH}} \tag{2.2.4.12}$$

Since x_i represents the fraction of protonated states for each titrating residue *i*:

$$\sum_i^N \langle x_i \rangle = \sum_i^N \left(1 - S_i^{unprot}\right) \tag{2.2.4.13}$$

$$\sum_i^N \left(1 - S_i^{unprot}\right) = \sum_j^N \frac{10^{pKa_j-pH}}{1+10^{pKa_j-pH}} \tag{2.2.4.14}$$

$$\sum_i^N S_i^{unprot} = \sum_j^N \left(1 - \frac{10^{pKa_j-pH}}{1+10^{pKa_j-pH}}\right) = \sum_j^N \left(\frac{1}{1+10^{pKa_j-pH}}\right) = \sum_j^N \left(\frac{1}{1+10^{-(pH-pKa_j)}}\right) \tag{2.2.4.15}$$

The DSR framework maps a set of *i* real sites to a set of *j* non-interacting quasi-sites. Assuming a one-to-one mapping of real to quasi sites (*i* = *j*), we obtain the following expression, which is the same expression as **equation 2.2.4.1**:

$$\sum_{i}^{N} S_i^{unprot} = \sum_{i}^{N}\left(\frac{1}{1+10^{-(pH-pKa_i)}}\right) \tag{2.2.4.16}$$

2.2.5. Reconstructing Apparent pK$_a$ from Microscopic pK$_a$

As the timescales of molecular dynamics simulation and complementary experimental methods for measuring pK$_a$ values differ on the order of several magnitude, for systems with considerable conformational flexibility, which transition between different conformation where the local electrostatic environment around the titrating residue can be different, the calculated microscopic pK$_a$ from CPHMD$^{MS\lambda D}$ simulations will not correspond to the experimentally measured macroscopic or apparent pK$_a$ value. In order to reconstruct the apparent pK$_a$ value from microscopic pK$_a$ values, the following relationship was derived.

Consider a system comprised of N conformational states, α = 1, N. Let those states have free energies G_α, yielding a distribution of population of each state ρ_α given by ρ_α α exp(-βG_α). Now, assume that each of these states is subject to a pH-dependent equilibrium over a set of titratable sites *i* = 1, m with pK$_a$ values of pK$_a^i$. The population of each state, α, will now depend on the external pH and the pK$_a$ values of each of the ionizable sites:

$$\rho_\alpha = \frac{\exp(-\beta G_\alpha)\left[1+\sum_{i=1}^{m}\exp\left(\ln 10\left(pK_{a,\alpha}^{i}-pH\right)\right)\right]}{\sum_\gamma \exp\left(-\beta G_\gamma\right)\left[1+\sum_{j=1}^{m}\exp\left(\ln 10\left(pK_{a,\gamma}^{j}-pH\right)\right)\right]}$$

$$\rho_\alpha = \frac{\exp(-\beta \Delta G_{0\alpha})\left[1+\sum_{i=1}^{m}\exp\left(\ln 10\left(pK_{a,\alpha}^{i}-pH\right)\right)\right]}{\sum_\gamma \exp\left(-\beta \Delta G_{0\gamma}\right)\left[1+\sum_{j=1}^{m}\exp\left(\ln 10\left(pK_{a,\gamma}^{j}-pH\right)\right)\right]}$$

where $\exp\left(\ln 10\left(pK_{a,\alpha}^{i} - pH\right)\right)$ represents the population of the unionized state at j in conformation γ at a given external pH.

To consider the fate of a single ionizable site, r, at a given external pH, we can derive the population of the unionized state as the summation of all conformational states in the unionized protonation state:

$$\rho^{r}(\text{unionized}) = \sum_{\alpha=1}^{N} \rho_{\alpha}^{r}(\text{unionized})$$

$$= \frac{\sum_{\alpha=1}^{N} \exp(-\beta \Delta G_{0\alpha}) \left[\exp\left(\ln 10\left(pK_{a,\alpha}^{r} - pH\right)\right)\right]}{\sum_{\gamma} \exp\left(-\beta \Delta G_{0\gamma}\right) \left[1 + \sum_{j=1}^{m} \exp\left(\ln 10\left(pK_{a,\gamma}^{j} - pH\right)\right)\right]}$$

and the population of the ionized states is given by:

$$\rho^{r}(\text{ionized}) = \sum_{\alpha=1}^{N} \rho_{\alpha}^{r}(\text{ionized})$$

$$= \frac{\sum_{\alpha=1}^{N} \exp(-\beta \Delta G_{0\alpha})}{\sum_{\gamma} \exp\left(-\beta \Delta G_{0\gamma}\right) \left[1 + \sum_{j=1}^{m} \exp\left(\ln 10\left(pK_{a,\gamma}^{j} - pH\right)\right)\right]}$$

To determine the apparent K_a of a given site r, which is defined as $[M^r][H^+]/[M^rH]$, where $[M^r]$ and $[M^rH]$ are the total concentrations of ionized and unionized states and $[H^+]$ is the proton concentration. If [M] represents the total concentration of the system M, then:

$$[M^{r}] = [M]\rho^{r}(\text{ionized}) \quad \textit{and} \quad [M^{r}H] = [M]\rho^{r}(\text{unionized})$$

From the definition of pH we also know:

$$pH = -\log_{10}[H^{+}]$$

$$[H^{+}] = 10^{-pH} = \exp(-\ln 10\, pH)$$

Therefore, the apparent K$_a$ is:

$$K_a^r(\text{apparent}) = \frac{[M]\rho^r(\text{ionized})[H^+]}{[M]\rho^r(\text{unionized})}$$

$$= \frac{[M]\sum_{\alpha=1}^{N} \exp(-\beta\Delta G_{0\alpha})\,[H^+]}{[M]\sum_{\alpha=1}^{N} \exp(-\beta\Delta G_{0\alpha})\left[\exp\left(\ln 10\left(pK_{a,\alpha}^r - pH\right)\right)\right]}$$

$$= \frac{[M]\sum_{\alpha=1}^{N} \exp(-\beta\Delta G_{0\alpha})\,[H^+]}{[M]\sum_{\alpha=1}^{N} \exp(-\beta\Delta G_{0\alpha})\exp\left(\ln 10\, pK_{a,\alpha}^r\right)\exp(-\ln 10\, pH)}$$

$$= \frac{[M]\sum_{\alpha=1}^{N} \exp(-\beta\Delta G_{0\alpha})\,[H^+]}{[M]\sum_{\alpha=1}^{N} \exp(-\beta\Delta G_{0\alpha})\exp\left(\ln 10\, pK_{a,\alpha}^r\right)[H^+]}$$

$$= \frac{\sum_{\alpha=1}^{N} \exp(-\beta\Delta G_{0\alpha})}{\sum_{\alpha=1}^{N} \exp(-\beta\Delta G_{0\alpha})\exp\left(\ln 10\, pK_{a,\alpha}^r\right)}$$

$$= \frac{\sum_{\alpha=1}^{N} \rho_\alpha}{\sum_{\alpha=1}^{N} \rho_\alpha \exp\left(\ln 10\, pK_{a,\alpha}^r\right)}$$

This expression in turn yields the result for apparent pK$_a$ expressed as equation **2.2.5.1**:

$$pK_a^r(\text{apparent}) = -\log_{10}\left(K_a^r(\text{apparent})\right)$$

$$= -\log_{10}\frac{\sum_{\alpha=1}^{N} \rho_\alpha}{\sum_{\alpha=1}^{N} \rho_\alpha \exp\left(\ln 10\, pK_{a,\alpha}^r\right)}$$

$$= \log_{10}\frac{\sum_{\alpha=1}^{N} \rho_\alpha \exp\left(\ln 10\, pK_{a,\alpha}^r\right)}{\sum_{\alpha=1}^{N} \rho_\alpha}$$

$$= \log_{10}\sum_{\alpha=1}^{N} x_\alpha \exp\left(\ln 10\, pK_{a,\alpha}^r\right) \qquad (2.2.5.1)$$

where x$_\alpha$ is the fractional population of conformational state α and $pK_{a,\alpha}^r$ is the pK$_a$ for the rth ionizable site in state α.

Chapter 3: Developing the Explicit Solvent CPHMD$^{MS\lambda D}$ Framework

3.1 Explicit Solvent CPHMD of Nucleic Acids in CHARMM

Note: Chapter 3.1 was adapted from the following references.[111,112]

3.1.1 New CHARMM Parameters for Protonated Adenine and Cytosine

Unlike pH-dependent protein activity, the role of pH regulation in nucleic acid systems has only been recently acknowledged.[6,7] As a result, there has been a lack of historical emphasis on the role of nucleobases in their alternative protonation states, and the development of the corresponding parameters have not been reported in either CHARMM or AMBER force fields. To estimate the charge distribution of protonated nucleotides for use with the CHARMM force field, we calculated the partial charges for the adenine and cytosine nucleobases in their neutral (unprotonated) and charged (protonated) states using the MMFF94 force field.[141] The difference in the partial charge was then added to the existing partial charge parameters for neutral adenine and cytosine in the CHARMM36 all-atom nucleic acid force field to assign the charge distribution for the protonated residues. A summary of the partial charge distribution and atom types for the protonated (calculated for this study) and unprotonated nucleic acids is reported in **Table 3.1.1.1**. Parameters for the bond, angle and dihedral energy terms for the protonated nucleic acid were adapted from existing nucleic acid parameters in CHARMM, and are reported in **Tables 3.1.1.2** through **3.1.1.4**. For the protonated adenine, the respective bonded parameters were obtained from guanine, specifically from the six-membered ring component that has atoms analogous to that of adenine (N1, H1, C2 and C6). For the protonated cytosine, the respective bonded parameters were obtained from a tautomeric form of neutral cytosine (obtained from patch *CYT1*).

Table 3.1.1.1: Charges and atom types assigned to the protonated and unprotonated states of titratable nucleic acids.

Name	Atom	Unprotonated		Protonated	
		Atom Type	Charge	Atom Type	Charge
ADE	H1	-	-	HN2	0.527
	N1	NN3A	-0.74	NN2G	-0.489
	C2	CN4	0.50	CN4	0.611
	C6	CN2	0.46	CN2	0.571
CYT	H3	-	-	HN2	0.52
	C5	CN3	-0.13	CN3	-0.174
	C2	CN1	0.52	CN1	0.75
	N3	NN3	-0.66	NN2C	-0.874
	C4	CN2	0.65	CN2	0.962
	N4	NN1	-0.75	NN1	-0.654
	H41	HN1	0.37	HN1	0.42
	H42	HN1	0.33	HN1	0.38

Table 3.1.1.2: Bond Parameters for Protonated Adenine and Cytosine. Parameters were adapted from comparable values from the tautomeric form of cytosine and guanine in CHARMM.

Atom Types	K_b	b_o
CN1 NN2C	350.0	1.335
CN2 NN2C	450.0	1.343
HN2 NN2C	474.0	1.01

Table 3.1.1.3: Angle Parameters for Protonated Adenine and Cytosine. Parameters were adapted from comparable values from the tautomeric form of cytosine and guanine in CHARMM.

Atom Types	K_θ	θ_o
CN4 NN2G CN2	70.0	131.1
NN2G CN4 NN3A	70.0	122.2
CN5 CN2 NN2G	70.0	107.8
NN2 CN1 NN2C	50.0	116.8
NN2C CN1 ON1C	130.0	123.8
CN1 NN2C HN2	37.0	121.2
CN1 NN2C CN2	85.0	119.1
CN2 NN2C HN2	37.0	121.2
CN3 CN2 NN2C	85.0	119.3
NN2C CN2 NN1	81.0	122.3

Table 3.1.1.4: Dihedral Parameters for Protonated Adenine and Cytosine. Parameters were adapted from comparable values from the tautomeric form of cytosine and guanine in CHARMM.

Atom Types	K_δ	n	δ
CN2 NN2G CN4 NN3A	0.2	2	180.0
HN2 NN2G CN4 NN3A	3.6	2	180.0
CN2 NN2G CN4 HN3	4.0	2	180.0
NN2C CN2 NN1 HN1	1.0	2	180.0
CN1 NN2C CN2 CN3	6.0	2	180.0
NN1 CN2 NN2C CN1	2.0	2	180.0
ON1C CN1 NN2C CN2	1.6	2	180.0
NN2 CN1 NN2C CN2	0.6	2	180.0
ON1C CN1 NN2C HN2	3.0	2	180.0
HN2 NN2C CN1 NN2	2.0	2	180.0
CN3 CN2 NN2C HN2	1.0	2	180.0
HN2 NN2C CN2 NN1	2.0	2	180.0

3.1.2 Optimization of Model Potential Parameters

The explicit solvent CPHMD$^{MS\lambda D}$ framework was implemented using the recently developed λ^{Nexp} functional form for λ in multi-site λ-dynamics (MSλD),[120,121] which is described in **Chapter 2**. To calibrate the external pH bath in our CPHMD$^{MS\lambda D}$ framework simulations, as with the previous implementation of CPHMD,[89-91] we used the calculated free energy of deprotonation ($\Delta G_{protonation}$) for each model compound, as the fixed biasing potential value in our simulation. In order to facilitate transitions between the two protonation states, we optimized the force constant (k_{bias}) on the variable biasing potential that was applied for each model compound.

It is interesting to note that without the application of the variable bias, no transitions between the protonated and unprotonated states were observed at conditions pH = pKa, where one should expect equal population of both states and the maximum transition rate between the two states (see **Figure 3.1.2.1**). At values of k_{bias} < 20 kcal/mol, there were very few transitions in λ phase space between the two states for the entire duration of a 1 ns trajectory. At values k_{bias} > 40 kcal/mol, transitions were rapid but the end states were not adequately sampled. The optimal value of k_{bias} for each nucleoside was selected by considering the competing needs for a high number of transitions and adequate sampling of the end-points (i.e., maintaining a high fraction of physical ligands (FPL) that were sampled). As illustrated in **Figure 3.1.2.2**, these two properties were observed to be anti-correlated to each other and there is a distinct range of k_{bias} values (between 25 and 35 kcal/mol) that yielded good transition rates and where more than 80% of the simulation is spent at the physically-relevant end-points. The optimized parameters for the model potentials are reported in **Table 3.1.2.1**.

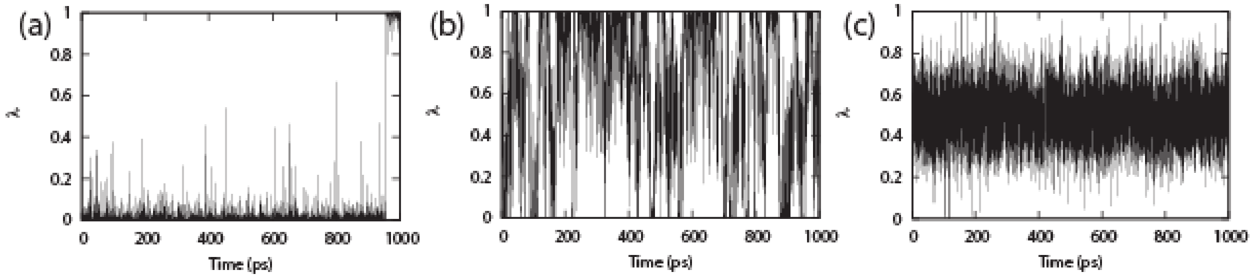


Figure 3.1.2.1: Transitions between the two protonation states of adenosine in λ phase space at pH = pK_a for a 1 ns trajectory with varying k_{bias} values of (a) 20, (b) 30 and (c) 40.

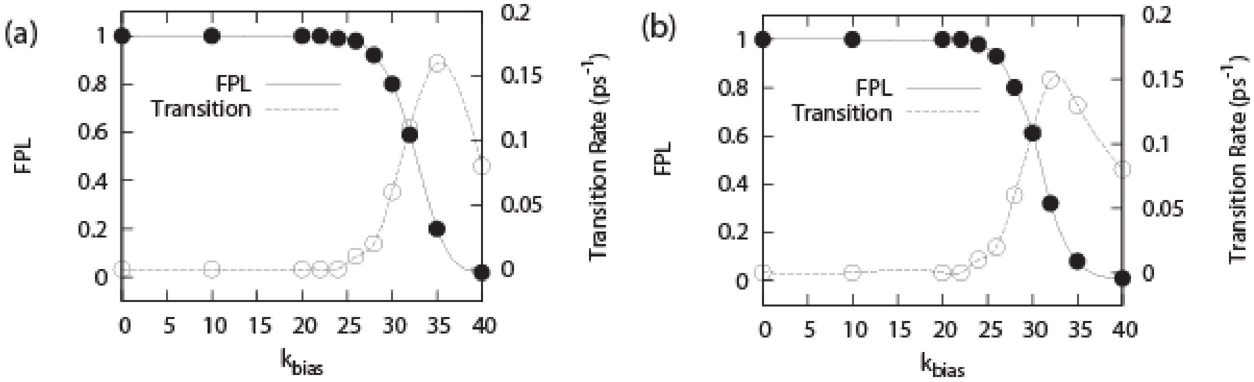


Figure 3.1.2.2: Effect of increasing k_{bias} on the transition rate and fraction physical ligand (FPL) for (a) adenosine and (b) cytosine. Sampling characteristics were obtained from 5 independent MD runs of 1 ns each.

Table 3.1.2.1: Parameters for the model potential.The reference pK_a is the experimental pK_a values for the model compounds.[142]

Nucleotide	$\Delta G_{protonation}$ (kcal/mol)	k_{bias} (kcal/mol)	Reference pKa
Adenine	19.39	29.75	3.50
Cytosine	75.24	27.75	4.08

The variable bias with a relatively large force constant of 28 to 30 kcal/mol that is required to achieve a reasonable number of transitions in our simulation may be rationalized by noting that the appearance of a full charge unit when titrating between the two states is likely to significantly perturb the solvent environment around the nitrogen atom. We suggest that time is required for the solvent to reorganize and fully accommodate the new charge distribution as the system titrates from the unprotonated to the protonated state. **Figure 3.1.2.3** provides a comparison of the radial distribution function (RDF) of water molecules surrounding the N1 atom of adenosine in its protonated and unprotonated state and indicates that considerable rearrangement of the first solvent shell upon ionization of the residue does occur. For the RDF that describes the distances between N1 and the TIP3P oxygen atoms, we observed that the charged protonated state had a first solvation shell (2.7 Å) that is slightly closer than the uncharged unprotonated state (2.9 Å). A more significant change, however, was observed for the RDF that describes the distances between N1 and the TIP3P hydrogen atoms in which the protonated state first solvation shell was pushed back (3.4 Å) compared to that of the unprotonated state (2.0 Å). These observations are consistent with the expectation that water molecules would orient their hydrogen atoms towards the partial negative charge of the nitrogen atom in the unprotonated state and subsequently would flip their hydrogen atoms “outwards” and orient their oxygen atoms closer towards the partial positive charge of the protonated hydrogen that is present in the protonated state. Similar trends were observed for the RDF of water molecules that surround the N3 atom of cytidine (data not shown). An analogous change in the RDF of water molecules around the protonated N5 atom of the substrate of dihydrofolate reductase was also observed with MD simulations that sampled different protein conformation that altered the water accessibility of the ligand pocket.[143]

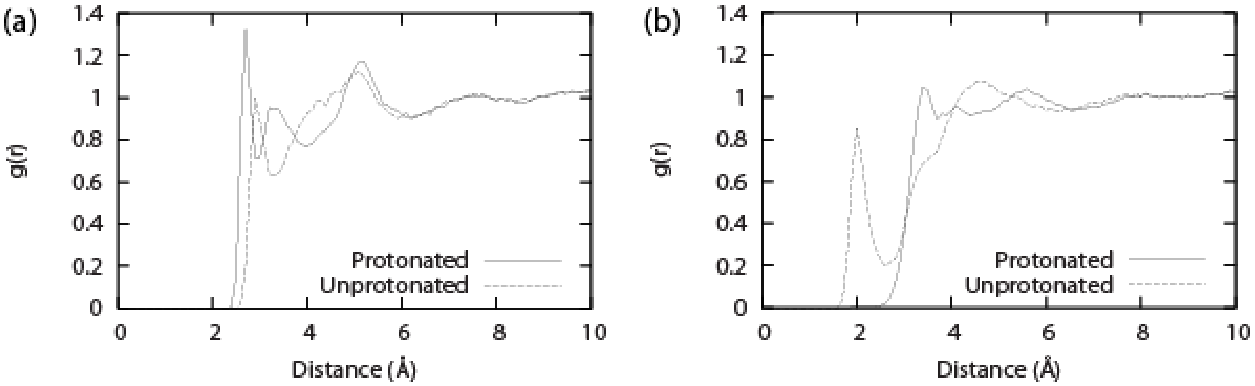


Figure 3.1.2.3: RDF of water molecules for (a) N1(ADE)-O(TIP3P) distances and (b) N1(ADE)-H(TIP3P) distances within a sphere of 10 Å from the N1 atom of adenosine in both protonated and unprotonated states.

3.1.3 Sampling Efficiency of Explicit Solvent CPHMD$^{MS\lambda D}$ simulations

The sampling efficiency as measured by the transition rates between the two protonation states in the explicit solvent CPHMD$^{MS\lambda D}$ framework is reasonably good with ~50 transitions per ns for our model compounds at pH = pKa. Given that the solvent reorganization upon the perturbation of a full charge unit was reported to be on a time scale of up to 3 ps in previous MD simulations[70] and that the mean duration of the physically-relevant protonation states in our simulations is 20 ps, the sampling characteristics of our system are sufficient to allow solvent reorganization to occur. However, the transition rate is markedly lower than what has been observed in CPHMD simulations that are performed using implicit solvent models.[89-91] It should be noted that our model potential parameters, specifically the k_{bias} values as implemented in the explicit solvent CPHMD$^{MS\lambda D}$ framework have been selected conservatively. For example, the transition rate can be doubled at the expense of reducing the FPL to 0.6 (**Figure 3.1.2.2**) which, provided that simulations are long enough to sufficiently enumerate the relative end-state populations, may be a better option for simulating larger-sized RNA systems where observing transitions between protonation states may be more challenging.

The more limited sampling efficiency of explicit solvent CPHMD simulations was also recently reported by Grubmüller and co-workers where the titration of an imidazole model compound achieved ~100 transitions in a 20 ns trajectory,[109] which is a rate of ~5 transitions per ns. Considering the computational expense of performing explicit solvent simulations, our rate of ~50 transitions per ns that is achieved with in our implementation of explicit solvent CPHMD is clearly advantageous. Finally, in **Table 3.1.3.1**, we present a comparison between the sampling characteristics of our simulation to that of previous work performed in the MSλD framework by Knight and Brooks for modeling series of inhibitors of HIV-1 reverse transcriptase.[122] Using the same force constant for the variable bias (i.e., k_{bias} = 7) as what was previously reported, we observed a significant drop in sampling performance with virtually no transitions observed between the two protonation states at pH = pKa. Our optimization of k_{bias} assisted in improving the sampling characteristics, but the transition rate still remains about four fold less efficient than previous work. We note that earlier work performed by Knight and Brooks modeled hybrid ligands in which the substituents did not differ significantly in terms of their partial charge distributions. Thus, the introduction of a full charge unit when titrating between the two states in the explicit solvent CPHMD$^{MS\lambda D}$ framework is likely to be the primary cause for the reduction sampling efficiency that we observe in the present simulations.

Table 3.1.3.1: Sampling characteristics of simulations performed at pH = pKa. Sampling characteristics of a two-state hybrid ligand in explicit water investigated in previous work (obtained from Table 3, hybrid ligand F).[122]

	Previous Work[a]	**Adenosine (Default)[b]**	**Cytidine (Default)[b]**	**Adenosine (Optimized)[c]**	**Cytidine (Optimized)[c]**
k_{bias}	7.00	7.00	7.00	29.75	27.75
FPL	0.780	1.000	1.000	0.828	0.832
Transitions (ps^{-1})	0.190	0.001	0.001	0.050	0.051

3.1.4 Convergence and Precision of Calculations

The challenges associated with sampling and convergence for CPHMD simulations has been reported on several occasions[78,91] and these are expected to be an even greater concern in explicit solvent CPHMD$^{MS\lambda D}$ where sampling efficiency is reduced. To validate the robustness of our explicit solvent CPHMD$^{MS\lambda D}$ framework in its ability to achieve adequate convergence, we performed a series of simulations at pH = pK$_a$ for our model compounds. The degree of convergence in our simulations was determined by calculating the unsigned deviation between the free energy of protonation, estimated from subsets of shortened trajectories, and the free energy of protonation that was estimated from ten 1ns trajectories. Different combinations of trajectory length and number of independent runs were systematically examined to determine the most cost effective tradeoff between computational expense and precision of the calculations. The results are summarized in **Figure 3.1.4.1** It was observed that individual trajectories required at least 100 ps to reliably observe any transitions between protonation states. In fact, we observed that a minimum simulation time of ~ 500 ps per trajectory was required to obtain a precision of ~0.20 kcal/mol in our calculations (**Figure 3.1.4.1a**) and running multiple shorter independent runs would not produce converged results unless the 500 ps threshold was crossed. Our results indicate that good precision can be achieved by using a total simulation time of 3 ns in the form of 3 independent runs of 1 ns each, where the unsigned deviations for the free energy of deprotonation was 0.05 kcal/mol for adenosine (**Figure 3.1.4.1b**). It should be noted that this level of precision was achieved in previous work three times more quickly for hybrid ligands whose charge distributions were similar.[122]

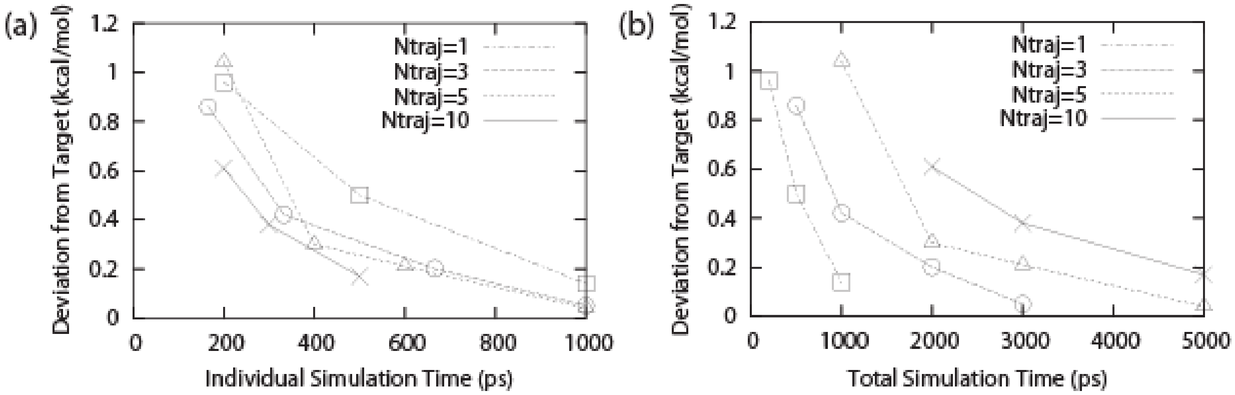


Figure 3.1.4.1: Unsigned deviation for the free energy of deprotonation of adenosine as a function of (a) total simulation time from all N trajectories and (b) individual simulation time of each of the N trajectories.

3.1.5 Calibration Curve of Model Systems: Adenosine and Cytidine

We calibrated our explicit solvent CPHMD$^{MS\lambda D}$ framework at 298 K at zero salt concentration. The reference pK$_a$ that was used in the calibration was the experimental pK$_a$ that was measured under similar conditions (25°C at zero ionic strength).[142] The titration curve of the model nucleoside compounds, adenosine and cytidine, are shown in **Figure 3.1.5.1**. The best-fit Henderson–Hasselbalch curve has a near ideal Hill coefficient for adenosine (n = 0.94) and cytidine (n = 0.93). The calculated pK$_a$ value of 3.50 for adenosine was in excellent agreement with experimental values and the pK$_a$ of 4.22 for cytidine is only slightly higher than the reference value by 0.14 pK$_a$ units. The accuracy of the calculated pK$_a$ values is determined primarily by the sampling efficiency at pH = pK$_a$ and the quality of the calibration of the $\Delta G_{protonation}$ values that are used to simulate distinct pH conditions. Our results demonstrate that a series of 3 x 1 ns simulations is sufficient to provide reasonably accurate results, which is significantly less than the 20 ns trajectory employed by Grubmüller and co-workers in their explicit solvent CPHMD implementation.[109]

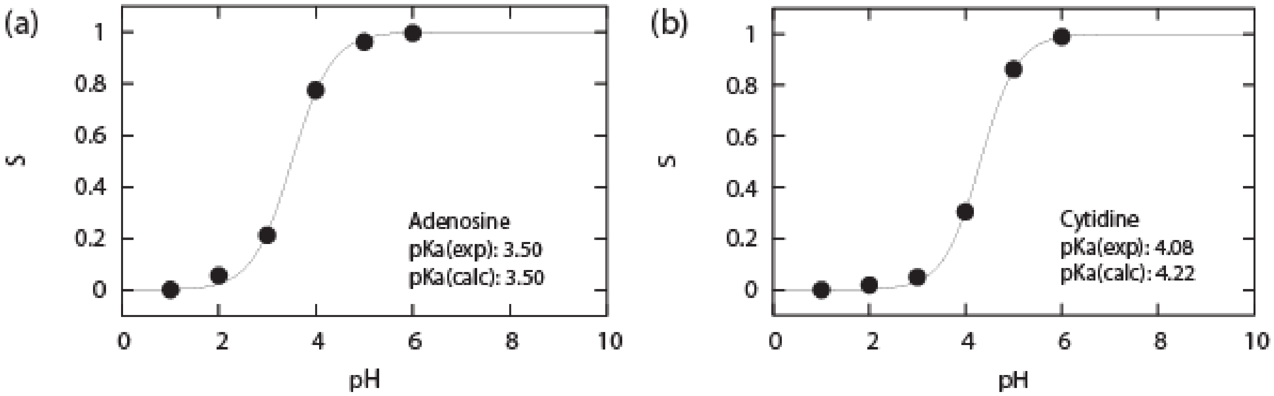


Figure 3.1.5.1: Sample titration curves for model nucleoside compounds, (a) adenosine and (b) cytidine.

Next, we tested our explicit solvent CPHMD$^{MS\lambda D}$ framework on single nucleotide test compounds, adenine monophosphate (AMP) and cytosine monophosphate (CMP), at zero ionic strength and the results are summarized in **Table 3.1.5.1**. The calculated pK_a values for AMP-5 and β-AMP-3 were 4.08 and 4.20 respectively. Compared to adenosine, the pK_a values of these nucleotide counterparts were slightly elevated by ~0.5 pK_a units. Similarly, the nucleotide counterparts of cytidine with pK_a values for CMP-5 and β-CMP-3 of 4.90 and 4.77, respectively, had slightly elevated pK_a values by ~0.5 pK_a units compared to cytidine. The calculated pK_a values for both 5'-phosphate and 3'-phosphate isomers of adenosine and cytosine are not statistically different at the 95% confidence interval. The increase in the calculated pK_a values from their nucleoside counterparts is expected, since the presence of the negative charge from the phosphate group may interact with the positively charged protonated base and weakly stabilize it, thus increasing the population of the protonated state and causing a corresponding increase in the calculated pK_a value.

In order to compare our calculated pK_a values with experimental results, we performed simulations that mimicked the ionic strength of the environment (i.e., 100-150mM NaCl) in which the experiments were performed.[144,145] By explicitly incorporating the salt environment, the calculated pK_a values are systematically lowered relative to those obtained from the zero ionic strength simulations. This shift in pK_a values is to be expected since the presence of Na^+ ions screens the electrostatic effects of the phosphate group. The results in **Table 3.1.5.1** indicate that our pK_a predictions had an average absolute error of 0.24 pK_a units compared to experiment and we conclude that our explicit solvent CPHMD framework is capable of making accurate quantitative predictions of pK_a values for simple nucleotides. These results also indicate that our model is capable of accounting for the differences between zero and non-zero ionic strength environments and highlights the importance of simulating the system at the appropriate ionic strength to mimic experimental conditions.

Table 3.1.5.1: Calculated and experimental pK_a values of test compounds.

Compound	[NaCl] (M)	Calculated	Experimental	Abs. Error
β-AMP-3	No salt	4.20 ± 0.06	-	-
β-AMP-3	0.15	3.79 ± 0.11	3.65	0.14
AMP-5	No salt	4.08 ± 0.03	-	-
AMP-5	0.15M	3.89 ± 0.16	3.74	0.15
β-CMP-3	No salt	4.77 ± 0.05	-	-
β-CMP-3	0.15M	4.56 ± 0.10	4.31	0.25
CMP-5	No salt	4.90 ± 0.07	-	-
CMP-5	0.10M	4.67 ± 0.08	4.24	0.43

3.1.6 Modeling Interactions between Adjacent Titrating Residues

Finally, we tested our explicit solvent CPHMD framework on dinucleotide sequences ADE-ADE, CYT-CYT and CYT-ADE at zero ionic strength, where both nucleotides were titrated simultaneously in the same simulation. The pK_a values were shifted upwards compared to the nucleoside model compounds for all sequences, ADE-ADE (4.08 ± 0.20 and 4.06 ± 0.16), CYT-CYT (4.93 ± 0.05 and 4.76 ± 0.09) and CYT-ADE (5.06 ± 0.07 and 3.85 ± 0.26), and were similar to the corresponding mononucleotide pK_a values. For some of the sets of pK_a calculations for the dinucleotide sequences, the Hill coefficient had more significant deviations from one compared to the monomeric compounds. Specifically, the value was lowered ($n < 0.8$) for 5 of the 9 sets of pK_a calculations. When the Hill coefficient deviates from one, it suggests that adjacent residues are interacting with each other in either a cooperative ($n > 1$) or anti-cooperative ($n < 1$) fashion. Cross-correlation analysis of the protonation states (data not shown) however, indicates only weakly correlated behavior, which suggests that the interaction between adjacent residues is not strong. The second set of pK_a calculations on CYT-ADE exhibited the lowest Hill coefficient ($n = 0.60$) indicating the strongest anti-cooperative behavior. Analysis of the individual titration curves as shown in **Figure 3.1.6.1** indicate that the S^{unprot} ratio shows the greatest deviation between the second set and the other two sets at pH 3. We analyzed the mean distance between the nitrogen atom that is protonated in CPHMD (i.e., N3 CYT and N1 ADE) of adjacent residues at pH 3 and the results are shown in **Figure 3.1.6.1**. In one simulation of the second set, the mean distance sampled was about 4 to 6 Å, in comparison to the typical values of 8 to 16 Å for all other simulations. We suggest that this simulation contributed significantly to the higher S^{unprot} ratio for the second set that in turn gave rise to the lower Hill coefficient. The lack of strong interactions between adjacent titrating residues in the other two sets of pK_a calculations of CYT-ADE is apparently due to the result of the lack of sampling of configuration space in which these two residues are close enough to influence each other's protonation state. We suggest that stronger cooperative or anti-cooperative effects are likely to be observed when modeling RNA structures with stable conformations in which the nucleobases are held in close proximity to one another.

Addendum: *Part of the analysis provided in this chapter was corrected in a later publication. See* ***Chapter 3.3*** *for the proper mathematical treatment of systems where the identity of adjacent residue cannot be distinguished (i.e. ADE-ADE and CYT-CYT dinucleotide).*

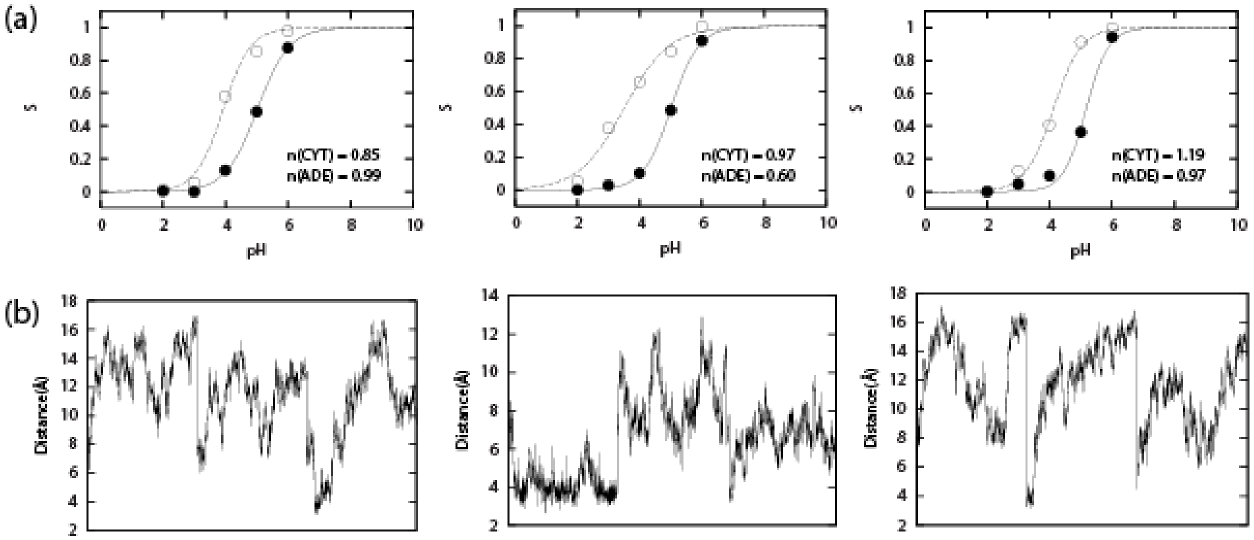


Figure 3.1.6.1: (a) Titration curves for CYT-ADE and (b) time series of distance between N3 CYT and N1 ADE atoms at pH 3 for all 3 sets of pK$_a$ calculation.

3.1.7 pK$_a$ calculations of a Model RNA using CPHMD$^{MS\lambda D}$ simulations

pH-dependent experimental observables, such as site-specific pK$_a$ values, may be used as an indicator of how accurately CPHMD$^{MS\lambda D}$ simulations reproduce pH-dependent properties. Unlike protein systems, where the site-specific pK$_a$ value of multiple ionizable residues for many proteins are readily available,[146] the literature of nucleic acid pK$_a$ research is much sparser with only a single pK$_a$ value measured for a handful of RNA systems. The lead-dependent ribozyme is, to the best of our knowledge, the most thoroughly-studied RNA system that has the largest number of experimentally-measured site-specific pK$_a$ values (**Figure 3.1.7.1a**).[147] Consequently, we have used it as a model system for benchmarking the performance of CPHMD$^{MS\lambda D}$ simulations in our work, and for understanding the potential challenges that one may encounter when modeling nucleic acid systems. The results from our calculations, as well as the appropriate comparisons with existing pK$_a$ values calculated using the NLPB equation[60] are summarized in **Figure 3.1.7.1**.

One of the key advantages of the CPHMD$^{MS\lambda D}$ framework is that no *a priori* information about the protonation states or the identity of the residues-of-interest of the system under investigation is required, as the local electrostatic microenvironment in conjunction with the external pH bath both serve to determine the protonation state at a given external pH. In our simulations, we simultaneously titrated all adenine and cytosine residues of the lead-dependent ribozyme. As summarized in **Figure 3.1.7.1b** and **Table 3.1.7.1**, we demonstrate good agreement with experimental pK_a values. Relative to experiments, our calculated pK_a values have an average unsigned error (AUE) of 1.3 pK_a units. With the exception of residue A16, the rank ordering of our calculated pK_a values also agree with experimentally measured values. The correlation coefficient between calculated and experimental pK_a value was 0.76, which is statistically significant at the 95% level. The precision of our calculated pK_a values, defined as the standard deviation of 3 independent sets of pK_a calculations was 0.3 pK_a units, which compares favorably to the average experimental uncertainty of 0.4 pK_a units. Our precision of 0.3 pK_a units translates to 0.4 kcal/mol, which is comparable to the precision of previous calculations on hydration free energy of benzene derivatives performed using MSλD.[122] The corresponding Hill coefficient of the calculated pK_a values were also generally below 1 (**Table 3.1.7.1**), suggesting that anti-cooperative interactions are the dominant mode in which titrating residues interact with one another.

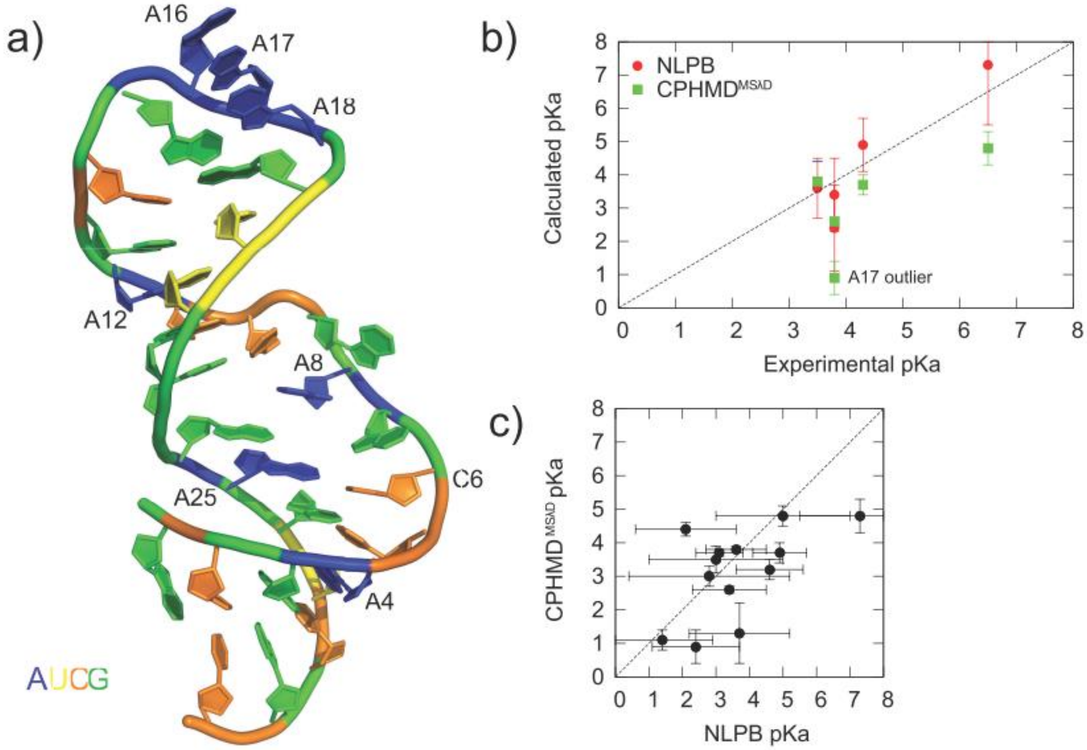


Figure 3.1.7.1: Comparison of pK_a values for lead-dependent ribozyme. (a) The lead-dependent ribozyme and the residues with experimentally measured pK_a values. (b) Correlation plot of calculated pK_a values from computational approaches (NLPB and CPHMD$^{MS\lambda D}$) and experimental pK_a values. (c) Correlation plot of pK_a values calculated from NLPB compared to CPHMD$^{MS\lambda D}$. The error bars denote the standard deviation of calculated pK_a values. All NLPB calculations were obtained from Honig, Pyle and co-workers.

Table 3.1.7.1: Comparison between experimental pK_a values with the calculated pK_a values obtained from CPHMD$^{MS\lambda D}$ simulations.

Residue	Exp. pK_a	CPHMD$^{MS\lambda D}$ Simulations		
		n	pK_a	Error
A4	< 3.0	0.4 ± 0.1	0.6 ± 0.1	-
A8	4.3 ± 0.3	0.7 ± 0.3	3.7 ± 0.3	-0.6
A12	< 3.0	1.1 ± 0.3	0.7 ± 0.3	-
A16	3.8 ± 0.4	0.7 ± 0.1	2.6 ± 0.1	-1.2
A17	3.8 ± 0.4	0.4 ± 0.0	0.9 ± 0.5	-2.9
A18	3.5 ± 0.6	0.6 ± 0.0	3.8 ± 0.1	0.3
A25	6.5 ± 0.1	0.4 ± 0.1	4.8 ± 0.5	-1.7
AUE				**1.3**
Precision			**± 0.3**	

Next, we ask if the shift in calculated pK_a values relative to the reference pK_a of the free unbound nucleobase is reasonable based on the structural considerations. A number of residues have been determined by experimental studies to be involved in Watson-Crick base pairing (indicated as "wc" in the structure column in **Table 3.1.7.2**). When adenine or cytosine participates in canonical base pairing as illustrated in **Figure 3.1.7.2**, their pK_a will be shifted lower relative to the reference value. This is because the nitrogen atoms (N1 for adenine, N3 for cytosine) that can be protonated serve as hydrogen bond acceptors, which make it energetically unfavorable for the base to be protonated. For all 9 residues in the lead-dependent ribozyme that are known to be base paired, $CPHMD^{MS\lambda D}$ predicted a lower pK_a relative to the reference compound. The exceptions are residues C2 and C30, which are located at the ends of the helix and are subject to fraying motions that weaken their base pairing interactions and increase their exposure to solvent. There is also a protonated $A25^{+}$•C6 base pair in the lead-dependent ribozyme, which is a configuration that raises the pK_a of the adenine base, as the protonated hydrogen on the N1 atom of adenine serves as a hydrogen bond donor to the N3 acceptor on cytosine (**Figure 3.1.7.2**). The calculated pK_a value of residue A25 was 4.8, which is shifted upwards from the reference value of 3.5 (**Table 3.1.7.2**). The calculated pK_a value of residue C6 was 1.8, which is shifted downwards from the reference value of 4.1 (**Table 3.1.7.2**). Thus, the direction of pK_a shifts of both residues in the $A25^{+}$•C6 base pair was correctly predicted. Lastly, based on the NMR data from Legault and Pardi, they reported that no cytosine residue in the lead-dependent ribozyme had an abnormally high pK_a value,[147] and our $CPHMD^{MS\lambda D}$ based calculations are consistent with their observations.

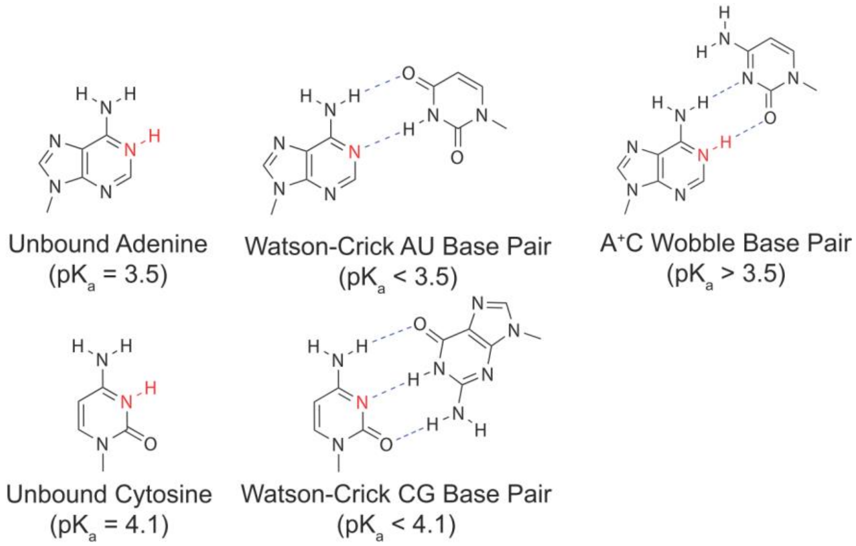


Figure 3.1.7.2: Illustration of adenine and cytosine and their hydrogen-bonding configuration in a canonical Watson-Crick base pair, and the protonated A$^+$•C base pair.

Table 3.1.7.2: Calculated pK$_a$ values of all adenine and cytosine residues in lead-dependent ribozyme obtained from NLPB calculations[60] and CPHMD$^{MS\lambda D}$ simulations indicate that both models produce consistent results and reasonable pK$_a$ shifts given structural considerations.

Residue	**Structure**	**NLPB**		**CPHMD$^{MS\lambda D}$**		**Abs Difference** (NLPB vs CPHMD$^{MS\lambda D}$)	**pK$_a$ shift** (wrt to ref pKa)
		pKa	**stdev**	**pKa**	**stdev**		
C2	wc	2.1 ±	1.5	4.4 ±	0.2	2.3	+
A4	wc	< 3.0		0.6 ±	0.1		-
C5	wc	3.0 ±	2.0	3.5 ±	0.4	0.5	-
C6	A$^+$C	2.8 ±	2.4	3.0 ±	0.3	0.2	-
A8		4.9 ±	0.8	3.7 ±	0.3	1.2	0
C10	wc	1.4 ±	1.5	1.1 ±	0.3	0.3	-
C11	wc	3.7 ±	1.5	1.3 ±	0.9	2.4	-
A12	wc	< 3.0		0.7 ±	0.3		-
C14	wc	4.6 ±	1.0	3.2 ±	0.3	1.4	-
A16		3.4 ±	1.1	2.6 ±	0.1	0.8	-
A17		2.4 ±	1.3	0.9 ±	0.5	1.5	-
A18		3.6 ±	0.9	3.8 ±	0.1	0.2	0
A25	A$^+$C	7.3 ±	1.8	4.8 ±	0.5	2.5	+
C28	wc	3.1 ±	0.7	3.7 ±	0.1	0.5	-
C30	wc	5.0 ±	2.0	4.8 ±	0.3	0.2	+
Average Unsigned Values			**1.5**		**0.3**	**1.1**	

3.1.8 Comparison to Implicit Solvent CPHMD Simulations

The accuracy of our pK$_a$ calculations compares favorably with established work on CPHMD simulations of proteins, which has a reported RMSE of 1.0 pK$_a$ units for surface-exposed residues and 1.5 pK_a units for buried residues.[91] The similar level of accuracy relative to established work on protein CPHMD simulation is encouraging, considering that constant pH simulations of nucleic acid systems in explicit solvent are met with several unique challenges. In nucleic acids, almost 50% of the residues present are titrating in unison and base-base interactions are extremely common given that they are the fundamental interactions that give rise to secondary and tertiary structure in RNA, analogous to how interactions between the amide backbone of protein contribute to protein secondary structure. This means that the probability of having coupled titrating interactions (i.e., when the protonation state of a nucleotide affects an adjacent residue and vice-versa) is high, which would increase the requirements for convergence.

The challenging nature of converging nucleic acid titrations is partially reflected in our longer 15 ns explicit solvent simulations, which is almost an order of magnitude longer than the shorter ~2ns simulations reported for protein CPHMD simulations.[91] However, the longer simulation time should be considered in the context that previous pK_a calculations on proteins which were performed in implicit solvent with temperature-replica exchange enhanced sampling,[91] where it is expected that more rapid sampling in both conformation space and titration coordinates would result in faster convergence. By contrast, our simulations were performed in explicit solvent and sampling of titration coordinates is slower due to the fact that the solvent needs to reorganize whenever the protonation state changes.[111] Despite the fact that implicit solvent models confer sampling advantages, there have been a number of unresolved issues based on earlier CPHMD work. For example, it has been reported that the Generalized-Born (GB) implicit solvent model underestimates the desolvation and buried charge-charge interactions which increases the error of predicted pK_a values of buried residues.[91] In addition, the approximations made in modeling hydrophobic interactions are known to cause structural compaction and possible distortion of the overall structure, which can be another source of error in pK_a calculations.[96,103] The above-mentioned sources of errors that are still unresolved in implicit solvent CPHMD are corrected with an explicit solvent representation of the protein's conformational dynamics,[103] highlighting the advantages of using an explicit solvent framework as we have done in our CPHMD$^{MS\lambda D}$ simulations. In addition, some RNA systems like the HDV ribozyme rely on specific Mg^{2+} ions to tune the local electrostatic environment around certain residues and consequently their pK_a values,[148] and the use of an explicit solvent model is needed to model this effect. Finally, it is worthwhile to consider that existing GB models used in earlier CPHMD simulations have been parameterized primarily against proteins,[73,74] and the naive application to nucleic acid systems is likely to introduce more errors if no re-paramaterization against nucleic acids is performed. Indeed, this expectation is consistent with earlier implicit solvent CPHMD simulations performed on the glmS ribozyme by Šponer, OtyepK_a and co-workers,[149] which demonstrated that implicit solvent models were unable to generate stable trajectories, and the simple Debye-Hückel screening function that is used to simulate the salt concentration appeared to have contributed to the inaccurate pK_a predictions.

3.1.9. Using CPHMD$^{MS\lambda D}$ Simulations to Investigate Localized pH-dependent Properties

The conventional approach in CPHMD simulations to investigate pH-dependent properties is to titrate the entire system. While this represents the most rigorously accurate approach, if one is investigating pH-dependent properties at a local site and the identity of titrating residues-of-interest are known, an informed choice to restrict the titration to a specific set of residues may be prudent. Such an approach would be justified, especially if available experimental data indicates that there are no other titrating residues in the vicinity of the local site within the pH range of interest that is being simulated. As an illustration of such informed

CPHMD$^{MS\lambda D}$ simulations, we performed a single-site titration of the lead-dependent ribozyme to investigate the A25^{+}•C6 base pair. In this single-site titration simulation, residue A25 is allowed to change its protonation state, but all other residues were assigned a protonation state that is consistent with their reference pK_a values (i.e., adenine and cytosine are unprotonated, guanine and uracil are protonated). From structural considerations, we know that the cytosine in the A25^{+}•C6 base pair will have a pK_a that is lower relative to its reference value. Thus, assigning it as a constitutively deprotonated residue (i.e., not titrating it in the CPHMD$^{MS\lambda D}$ simulations) is a well-justified approximation. The resulting pK_a value from single-site CPHMD$^{MS\lambda D}$ simulations is 6.1, which is close to the experimental value of 6.5. Lastly, from NMR studies we know that the pK_a value of residue A25 decreases from 6.5 to 5.9 when the salt concentration is increased from 100mM and 500mM NaCl, due to the additional screening effect in a higher ionic strength environment.[150] Using single-site CPHMD$^{MS\lambda D}$ simulations, our calculated pK_a values decreased from 6.1 to 5.0 when the simulated salt concentration was increased from 100mM to 500mM. The pK_a calculations agree well with experiment, highlighting that CPHMD$^{MS\lambda D}$ simulations can be used to model the effects that ionic strength has on the protonation state of residues in RNA structures.

3.1.10 Conformational Dynamics and Coupled Titrating Interactions

The interplay between conformational dynamics and protonation states, the process of how local structural changes modify the electrostatic microenvironment around residues to cause a change in protonation state, is well documented in many RNA systems. Some examples include retrovirus pseudoknot structures,[26] the intramolecular stem-loop of the spliceosome complex,[35] the peptidyl-transferase center of the ribosome,[27-32] and helix 69 of the 50S ribosomal subunit,[33,34] where the pH-dependent dynamics of these RNA complexes are known to alter their structure and function. Similar observations have also been reported in proteins as well.[78,123,151] Thus, the importance of conformational dynamics in RNA systems in influencing protonation states, together with the high possibility of coupled titrating interactions due to the ubiquitous nature of base-base interactions, is going to be of emerging interest in the field of CPHMD simulations of nucleic acids.

The GAAA tetraloop of the lead-dependent ribozyme is a conformationally dynamic motif common to many RNA structures.[152] It contains three titratable adenine residues. It serves as an excellent model for examining the interplay between conformational dynamics and coupled titrating interactions in our CPHMD$^{MS\lambda D}$ simulations. The lowest energy conformation as determined by NMR spectroscopy is one where the three adenine residues (A16, A17, A18) adopt a triply stacked conformation as shown in **Figure 3.1.10.1a**. Considering the close proximity of these residues, it is likely that their protonation states are coupled. Examination of the distance between the N1 atoms at pH 2 indicates that A17 and A18 remain stacked on top of each other and they do not move more than 4 Å away for most of the simulation as indicated in **Figure 3.1.10.1**. This distance is much lower than the 6 Å distance that we previously reported in dinucleotides, which is the range where only weak interactions between adjacent nucleotides were observed.[111] This suggests that there may be anti-cooperative interactions between the two residues, which was confirmed by the near perfect correlation between the unprotonated state of A17 and the protonated state of A18 when the N1-N1 interatomic distance is less than 4 Å (**Figure 3.1.10.1**). In one of the MD runs (highlighted in grey), this distance increased to 15 Å, which was concomitant with A17 transitioning to and maintaining a predominantly protonated state ($\lambda_{A17,unprotonated} = 0$). In this run, the lead-dependent ribozyme sampled and remained trapped in an alternative unstacked conformation as illustrated in **Figure 3.1.10.1b**, which altered the electrostatic microenvironment around each residue and consequently their protonation states. At the same time, we observed a loss in correlation between the protonation states of A17 and A18 whenever their distance exceeded 6 Å. The results presented here are the first example of a microscopic examination of the interplay between local conformational dynamics and coupled titrating interactions in nucleic acid literature.

This physically realistic model of coupled titrating interactions that respond to conformational dynamics in our CPHMD$^{MS\lambda D}$ simulations helps account for the calculated pK_a value of 0.9 of residue A17. Since A17 and A18 remain stacked for most of the simulation, the positive electrostatic environment generated by A18 would artificially depress the tendency of A17 to achieve protonation. In addition, apart from the strong correlations between the protonation states of these 2 residues, the pK_a value calculated from titrating residue A17 only, with A18 permanently assigned to its protonated state was less than 1, confirming that the coupled A17-A18 interactions are responsible for the shifted pK_a value of residue A17.

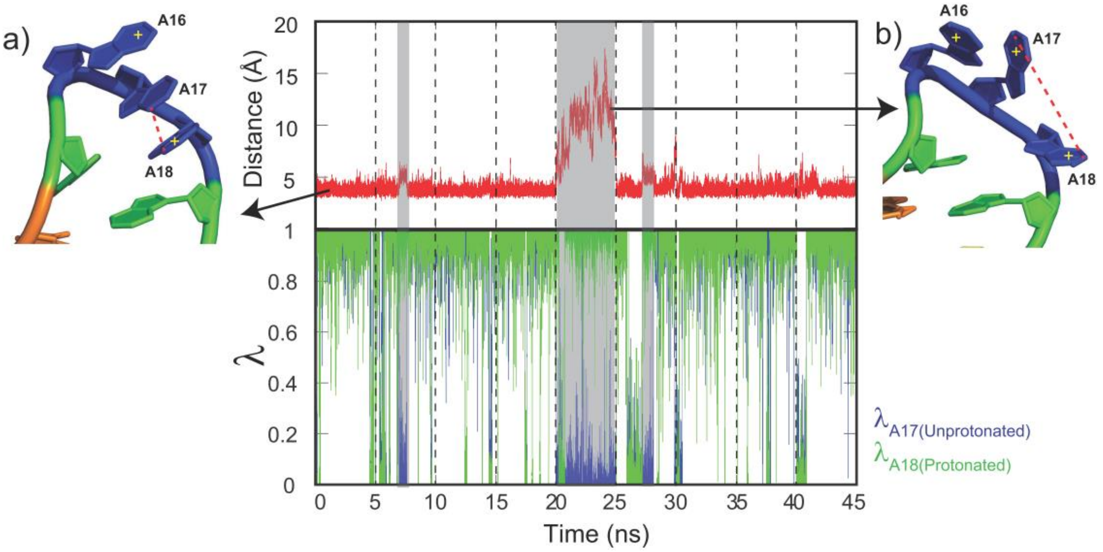


Figure 3.1.10.1: Concatenated trajectories from 9 independent 5 ns runs that describe (top) the distance between N1 atoms of A17/A18 at pH 2 and (bottom) the corresponding λ_{unprot} state of A17 and λ_{prot} state of A18. (a) A typical triply-stacked lowest energy conformation maintained throughout most of the simulations, (b) an alternative unstacked conformation that resulted in a decoupling of A17-A18 interactions.

To reconcile the difference between calculated and experimental pK_a values of A17, we suggest that the coupled titrating interactions observed in our simulations become decoupled on the longer NMR timescale, when the loop undergoes multiple excursions between various alternate conformations. This would be consistent with the work of Pardi and co-workers which indicated that the GAAA tetraloop of the lead-dependent ribozyme adopts at least one other low energy conformation.[153,154] The existence of alternative conformations sampled by such GNRA tetraloops has been suggested by fluorescence spectroscopy experiments,[155] and possible alternative conformations was suggested based upon temperature replica exchange MD simulations.[156] A visual inspection of these alternate conformations indicate that more than half of them are different from the triply-stacked conformation, which suggests that the sampling limitations in straightforward MD simulations that prevents us from accessing the other alternative conformations sampled may be responsible for the reduced accuracy of residue A17's pK_a value.

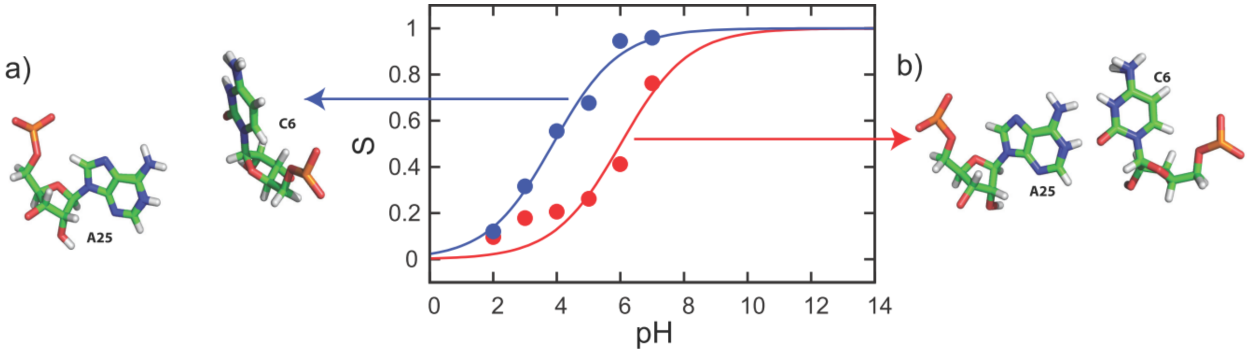


Fig. 3.1.10.2: Titration curves from reprocessed trajectory that maintained a (red) closed conformation and (blue) semi-open conformation resulted in distinct pK_a values of 6.0 and 3.9 respectively. (a) A sample snapshot of a semi-open conformation and (b) a typical closed conformation.

The A25$^+$•C6 base pair in the lead-dependent ribozyme is another interesting case where we can examine the effects of coupled titrating interactions in our CPHMD$^{MS\lambda D}$ simulations. Examination of the hydrogen bonding distances indicates that the system oscillates between closed (**Figure 3.1.10.2b**) and semi-open (**Figure 3.1.10.2a**) conformations. The closed conformation is consistent with the NMR structure of the A25$^+$•C6 base pair, and it promotes the protonation of A25. The semi-open state on the other hand exposes A25 to the solvent and is therefore expected to have a protonation equilibrium that is similar to the reference adenosine compound. When we reprocessed the concatenated trajectory from all simulation runs and extracted the segments that maintained a proper A25$^+$•C6 geometry, the resulting pK_a was 6.0 as shown in **Figure 3.1.10.2**. Conversely, the pK_a calculated from segments of the trajectory that did not maintain the base-paired geometry was 3.9, which is close to the reference pK_a of 3.5 for a solvent-exposed adenosine. Thus, the excursions between these two conformations accounts for the calculated pK_a value of 4.8 for residue A25, which is lower than the experimental pK_a of 6.5. The A25$^+$•C6 base pair as we have seen, has similar conformational sampling challenges as the GAAA tetraloop. The larger underprediction of 1.7 pK_a units that corresponds to a free energy difference of 2.3 kcal/mol, is thus consistent with observations in the literature for free energy calculations in systems with higher demands in terms of conformational sampling usually have a lower accuracy as compared to systems that exhibit lesser conformational dynamics.[157,158]

3.1.11 Conclusion: The First Viable Explicit Solvent CPHMD$^{MS\lambda D}$ Simulation Framework was Developed

In this chapter, we reported the first implementation of an explicit solvent CPHMD framework for nucleic acids. By adopting the new functional form λ^{Nexp} for λ that was recently developed for multi-site λ-Dynamics (MSλD), we demonstrated good sampling characteristics in which rapid and frequent transitions between the protonated and unprotonated states at $pH = pK_a$ are achieved, while sampling the physically-relevant protonation states for more than 80% of the trajectory. Compared to existing implementations of explicit solvent CPHMD, the sampling in our method sees a 10-fold improvement, while maintaining sufficient residency time of the physical protonation states to ensure proper solvent reorganization. pK_a values calculated for simple nucleotides are in a good agreement with experimentally measured values with a mean absolute error of 0.24 pK_a units, affirming that our explicit solvent $CPHMD^{MS\lambda D}$ framework has the ability to make accurate quantitative predictions for simple nucleotide systems. This work was followed by the first demonstration of an explicit solvent $CPHMD^{MS\lambda D}$ simulation of a complex RNA structure, the lead-dependent ribozyme. Our initial pK_a values calculated from $CPHMD^{MS\lambda D}$ simulations agree well with experimental pK_a values with an average unsigned error of 1.3 pK_a units. The accuracy of our pK_a calculations is comparable to established CPHMD work in proteins and the direction of the pK_a shifts for all residues in the lead-dependent ribozyme are also accurately predicted when compared to experimental data or structural considerations. Using the GAAA tetraloop and the $A^+\bullet C$ base pair of the lead-dependent ribozyme as model systems, we demonstrated that $CPHMD^{MS\lambda D}$ simulations are able to model the effects that conformational dynamics and coupled titrating interactions have on the protonation equilibria of titrating residues. Therefore, this work paves the way for the utilization of $CPHMD^{MS\lambda D}$ simulations as a tool to investigate pH-dependent biological properties of RNA macromolecules.

3.2 Sampling Challenges in Explicit Solvent CPHMD

Note: Chapter 3.2 was adapted from the following references.[113]

3.2.1 Sampling Improvements with pH-based Replica Exchange

Our earlier work on the lead-dependent ribozyme in **Chapter 3.1** has suggested that conformational sampling challenges may be responsible for some of the outliers in our CPHMD$^{MS\lambda D}$ simulations. To address this concern, we implemented a replica exchange protocol in pH space, and tested its performance in terms of sampling efficiency of titration coordinates and general accuracy of predicted pK_a values. The microscopic pK_a values calculated from pH-REX simulations, as summarized in **Table 3.2.1.1**, are consistent with previous work that utilized CPHMD$^{MS\lambda D}$ with conventional MD simulations.[112] As illustrated in **Figure 3.2.1.1**, up to an 8-fold improvement in the transition rates in λ-space is observed in our pH-REX simulations. The sampling improvement of titration coordinates results in faster convergence, which is demonstrated by fact that pH-REX sampling achieves the same level of accuracy using a total simulation time that is 5-fold shorter than conventional CPHMD$^{MS\lambda D}$ simulations. In addition, we also observe that the improvement in λ-space sampling for the residues of the lead-dependent ribozyme is higher than that of the 3-fold improvement in single nucleotide compounds.

In complex RNA structures, where multiple residues are titrated simultaneously, the coupled interactions between these titrating groups lead to slower convergence, because the sampling of titration coordinates is hindered by the protonation states of adjacent interacting titrating groups.[112] The variable biases applied in conventional CPHMD$^{MS\lambda D}$ simulations only serve to flatten the potential energy surface of each λ variable, but the orthogonal barriers that arise from these coupled titrating interactions are not addressed. Unlike the recent methodological advances in enhanced sampling strategies reported by Yang and co-workers,[151,159] pH-REX sampling does not directly address these orthogonal barriers *per se*. However, it does periodically shuffle conformations to a higher or lower pH where all residues adopt a uniform protonation state. We suggest that this process effectively decouples the protonation states of interacting residues, allowing one to ameliorate the sampling issues related to these orthogonal barriers.

Table 3.2.1.1: Calculated pK$_a$ values from conventional and pH-REX CPHMD$^{MS\lambda D}$ simulations of the lead-dependent ribozyme demonstrate a similar level of accuracy.

Residue	**Exp pK$_a$**	**Conventional CPHMD$^{MS\lambda D}$ (3x5ns)**			**pH-REX CPHMD$^{MS\lambda D}$ (3ns)**		
		n	**pK$_a$**	**Error**	**n**	**pK$_a$**	**Error**
A4	<3.1	0.4 ± 0.1	0.6 ± 0.1		1.3 ± 0.5	0.9 ± 0.4	
A8	4.3 ± 0.3	0.7 ± 0.3	3.7 ± 0.3	-0.6	0.9 ± 0.4	3.8 ± 0.6	-0.5
A12	<3.1	1.1 ± 0.3	0.7 ± 0.3		1.0 ± 0.3	0.6 ± 0.2	
A16	3.8 ± 0.4	0.7 ± 0.1	2.6 ± 0.1	-1.2	0.7 ± 0.1	2.6 ± 0.0	-1.2
A17	3.8 ± 0.4	0.4 ± 0.0	0.9 ± 0.5	-2.9	1.0 ± 0.6	1.1 ± 0.5	-2.7
A18	3.5 ± 0.6	0.6 ± 0.0	3.8 ± 0.1	0.3	0.8 ± 0.1	3.6 ± 0.0	0.1
A25	6.5 ± 0.1	0.4 ± 0.1	4.8 ± 0.5	-1.7	0.5 ± 0.1	4.5 ± 0.2	-2.0
AUE				**1.3**			**1.3**

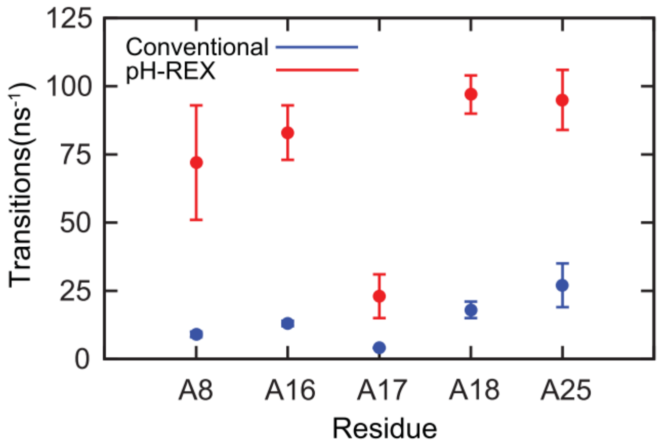


Figure 3.2.1.1: pH-REX CPHMD$^{MS\lambda D}$ simulations accelerates sampling of titration coordinates by up to 8-fold in the lead-dependent ribozyme.

Having demonstrated that pH-REX accelerates sampling of titration coordinates, we now explore the apparent underprediction of the pK_a value of residue A17, which is situated in the GAAA tetraloop of the lead-dependent ribozyme. We performed an initial 15 ns simulation of the excised GAAA tetraloop for pH values between 1 to 4, and compared the results to conventional CPHMD$^{\mathrm{MS\lambda D}}$ simulations. As summarized in **Table 3.2.1.2**, the calculated pK_a of residue A17 from the conventional simulations is 1.4, compared to the pK_a of 2.3 obtained using pH-REX sampling. Extending our simulations for an additional 15 ns confirmed that the pK_a value has converged. On the whole, pH-REX sampling yields systematic improvement of the predicted pK_a values of the GAAA tetraloop, where the average unsigned error (AUE) was reduced to 0.7 pK_a units, which is 50% lower than our previous work.[112]

Table 3.2.1.2: pH-REX CPHMD$^{\mathrm{MS\lambda D}}$ simulations of the GAAA tetraloop of lead-dependent ribozyme improved the accuracy of calculated pK_a values compared to straightforward MD simulations.

Residue	**GAAA**				**AAA**
	Exp pK_a	**Conventional CPHMD$^{\mathrm{MS\lambda D}}$ (0-15ns)**	**pH-REX CPHMD$^{\mathrm{MS\lambda D}}$ (0-15ns)**	**pH-REX CPHMD$^{\mathrm{MS\lambda D}}$ (16-30ns)**	**pH-REX CPHMD$^{\mathrm{MS\lambda D}}$ (0-15ns)**
A16	3.8±0.4	3.2 ± 0.2	3.1 ± 0.1	3.3 ± 0.1	3.5 ± 0.1
A17	3.8±0.4	1.4 ± 0.3	2.3 ± 0.6	2.6 ± 0.4	3.5 ± 0.1
A18	3.5±0.6	3.9 ± 0.1	3.9 ± 0.1	4.0 ± 0.1	3.9 ± 0.1
AUE		1.1	0.9	0.7	

The apparent underprediction of the pK_a value of residue A17 originates from the anti-cooperative interactions between residues A17 and A18, which artificially suppresses the ability of A17 to adopt the protonated state at low pH conditions.[112] This arises from the triply stacked conformation of the GAAA tetraloop, which is characterized by short interatomic distances between the N1 atoms of the two residues. We analyzed this interatomic distance in our simulations of the GAAA tetraloop at pH 1, in the context of another reference simulation of the AAA trinucleotide sequence, which has no structural elements imposing conformational restrictions on it. As shown in **Figure 3.2.1.2**, the N1-N1 distances sampled in our pH-REX simulations of the GAAA tetraloop ranged from 2 to 10 Å, which are intermediate between the conventional GAAA tetraloop simulations (2 to 6 Å) and the AAA trinucleotide simulations (6 to 18 Å). The conformational space sampled using pH-REX is reasonable as it does not exhibit more dynamical behavior than the free AAA trinucleotide, but it also samples more conformations than conventional CPHMD$^{MS\lambda D}$ simulations of the GAAA tetraloop. This trend of exploring progressively larger N1-N1 distances results in more weakly coupled interactions that is reflected in the pK$_a$ value of the central adenine residue, which increases from 1.4 to 2.3 to 3.5 (**Table 3.2.1.2**). In addition, we also observed a slight difference in the distribution of the N1-N1 distances between the first 15 ns and the subsequent 15 ns trajectory of the excised GAAA tetraloop. Specifically, the "close contact" region of 3 to 6 Å that denotes the initial stacked conformation was partially populated in the first 15 ns, which suggests that the system was still equilibrating for part of that trajectory. This suggests that sufficient equilibration on the order of a few ns may be required, and metrics such as RMSD relative to the initial structure may be used to determine when equilibration is complete, particularly if one is expecting a significant conformational change in an alternate pH environment.

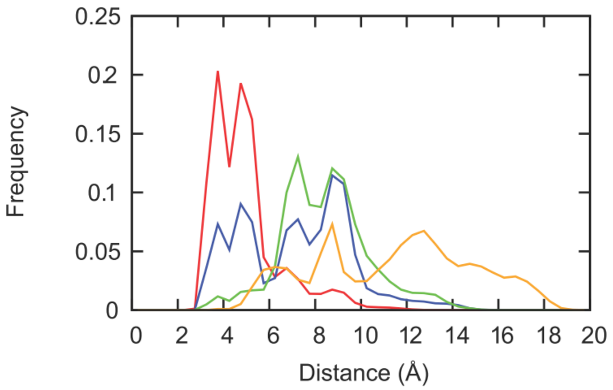


Figure 3.2.1.2: Distribution of the interatomic distance of the N1 atoms of residue A17 and A18 of the GAAA tetraloop at pH 1 for a conventional 15ns MD simulation (red trace), the first 15 ns pH-REX simulation (blue trace) and the next 15 ns pH-REX simulation (green trace), compared to a 15 ns pH-REX simulation of the AAA trinucleotide (orange trace).

The experimentally measured pK_a values are a superposition of the microscopic pK_a values of the various conformations visited by the GAAA tetraloop on the timescale accessible to NMR measurements,[112] and the various pH values at which such measurements were recorded. We clustered the conformations sampled by pH-REX simulations at pH 1 (low pH) and 4 (high pH), and the representative structures are illustrated in **Figure 3.2.1.3** The initial triply-stacked conformation (**Figure 3.2.1.3b**), which is representative of the NMR structure solved at physiological pH is known to lower the pK_a value of residue A17. While it may be the dominant conformation sampled at high pH, this conformation is populated only 20% of the time at low pH. We observe that the dominant conformation sampled at low pH is partially unstacked, where the N1-N1 distance is increased to 9.3 Å (**Figure 3.2.1.3a**), and a fully unstacked conformation is observed 10% of the time (**Figure 3.2.1.3c**). Interestingly, these unstacked conformations are populated 21% of the time at higher pH. The significant improvement in our pK_a predictions for residues in the GAAA tetraloop correspond to the sampling of these alternative conformations, suggesting that sampling using pH-REX provides a more accurate model of the tetraloop's pH-dependent dynamics. Lastly, our results also indicate that pH-REX CPHMD$^{MS\lambda D}$ simulations can be used to identify the dominant conformation of nucleic acid systems in different pH environments or low population conformations at physiological pH. With the discovery of pH-dependent transient states in both RNA and DNA systems that have been suggested to be functionally important,[49,50] we anticipate that pH-REX CPHMD$^{MS\lambda D}$ simulations will provide further structural and mechanistic insight into the findings gleaned from experimental studies, especially in situations where direct experimental characterization of such transient states are challenging.

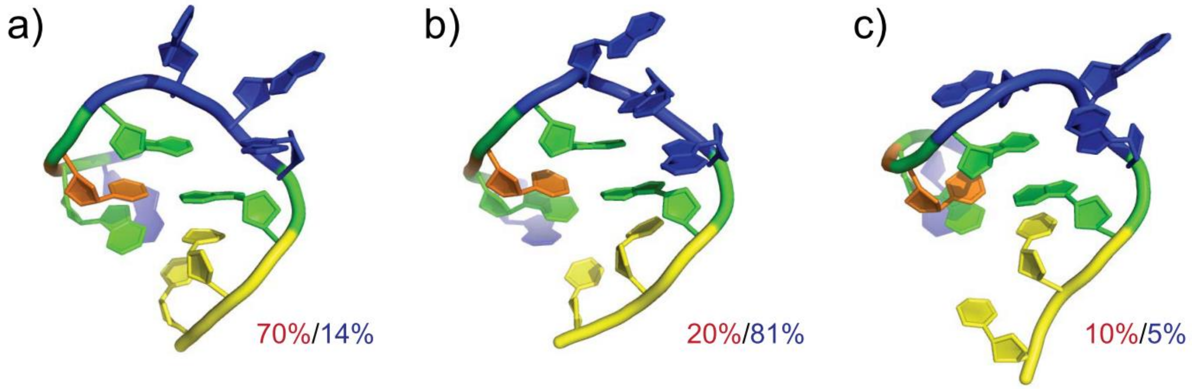


Figure 3.2.1.3: Representative conformations from a cluster analysis of the pH-REX trajectory of the GAAA tetraloop and their relative populations sampled at pH 1 (in red) and pH 4 (in blue).

3.2.2. Scalability of Explicit solvent pH-based Replica Exchange CPHMD$^{MS\lambda D}$ simulations

Thus far, we have shown that pH-REX CPHMD$^{MS\lambda D}$ simulations are effective in modeling accurate pH-dependent dynamics of small nucleic acid motifs like the GAAA tetraloop. We now extend this to demonstrate the scalability of pH-REX CPHMD$^{MS\lambda D}$ simulations to larger systems. Our initial pK_a calculations on the full-length ribozyme (**Table 3.2.1.1**) yielded similar results to conventional CPHMD$^{MS\lambda D}$ simulations, which suggest that the sampling efficiency is not as high as in our simulations of the excised GAAA tetraloop. Since conformational diffusion across pH space is responsible for enhancing sampling, increasing the total number of accepted Monte Carlo (MC) moves should improve the accuracy of calculated pK_a values. In the full-length ribozyme, the majority (i.e., 10 out of 15) of the residues are base paired and have pK_a values of less than 3. Thus, unlike high pH conditions where most of the titrating residues adopt a uniform protonation state, at low pH conditions the majority of residues would be titrating and a more pronounced potential energy difference between adjacent replicas, and consequently lower MC exchange rate is expected. The MC exchange rate of the excised tetraloop was at least 30% at low pH conditions, which is 3 times higher than that of the full-length ribozyme. This lower exchange rate in the full-length ribozyme is correlated with a reduction in the sampling of titration coordinates, particularly for residue A17 (**Figure 3.2.2.1**). As shown in **Figure 3.2.2.1a**, increasing the frequency of MC attempts from every 1.0 ps to 0.1 ps increased λ-transitions to ~350-650 ns^{-1}. While the significant improvement in the sampling of titration coordinates is encouraging, we note that the variation in transition rates between the 3 independent runs is significant. This could reduce the reproducibility of computed results, especially if the conformation space sampled by different independent runs is not uniform due to the disparate sampling in titration coordinates. Furthermore, prior work by Baptista and co-workers has shown that re-equilibration of the solvent induced by the introduction of a charged protonation state requires 1 to 3 ps,[70] and we have also seen a similar solvent reorganization triggered by a protonation state change in explicit solvent $\text{CPHMD}^{\text{MS}\lambda\text{D}}$ simulation.[111] Using the mean solvent relaxation time of 2 ps, a conservative estimate for the maximum transition rate is ~500 ns^{-1}, and in some instances, such as residue A8, the transition rate was above this value. While we acknowledge that the pH-REX protocol maintains detailed balance and the results should *in principle* be unaffected by the MC exchange frequency, it is possible that too frequent successful exchanges between replicas may not allow for sufficient solvent relaxation to occur, and this could lead to non-ergodic behavior.

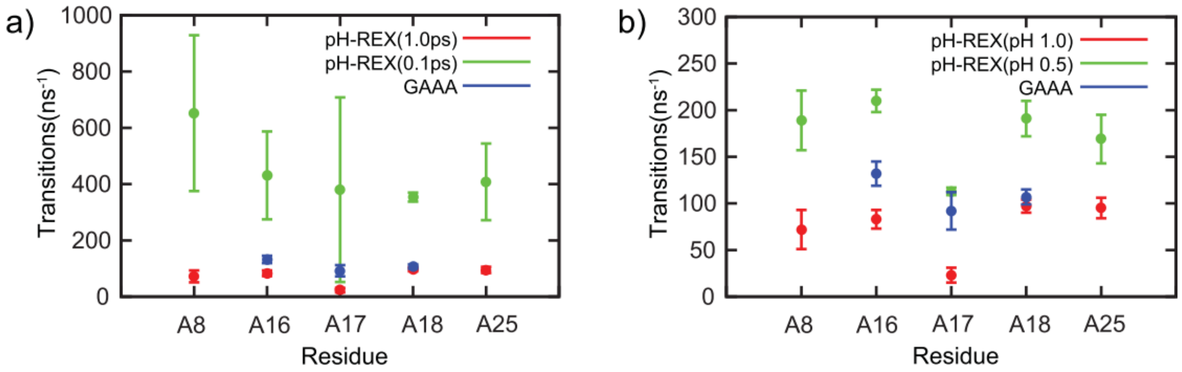


Figure 3.2.2.1: Effects on titration coordinates sampling by (a) increasing the MC attempt frequency and (b) reducing pH window spacing from 1.0 to 0.5.

Table 3.2.2.1: pH-REX CPHMD$^{MS\lambda D}$ simulations of the full-length lead-dependent ribozyme at 0.5 pH window spacing demonstrate comparable results to the GAAA tetraloop within 13 ns.

Residue	**Exp pK$_a$**	**pH-REX CPHMD$^{MS\lambda D}$**				
		(0-3ns)	**(3-8ns)**	**(8-13ns)**	**(13-18ns)**	**(18-23ns)**
A16	3.8±0.4	2.6 ± 0.1	2.7 ± 0.2	2.9 ± 0.2	2.9 ± 0.1	2.8 ± 0.1
A17	3.8±0.4	1.4 ± 0.3	1.5 ± 0.6	1.8 ± 0.6	2.4 ± 0.1	2.4 ± 0.1
A18	3.5±0.6	3.8 ± 0.1	3.8 ± 0.1	3.9 ± 0.1	3.8 ± 0.1	3.7 ± 0.0

Our observations on increasing the MC exchange frequency differs with the findings reported by Roitberg and co-workers, where no performance degradation was observed at higher MC exchange frequencies.[80] This difference is likely due to the fact that our model uses an explicit solvent representation where solvent reorganization needs to be accounted for, whereas the work of Roitberg and co-workers was performed in implicit solvent, which adiabatically adjusts to the protein conformation. Instead of attempting to increase the MC exchange frequency, one may also increase the probability of exchange by reducing the potential energy difference between adjacent windows (i.e. reduce the pH-spacing). In simulations using a smaller spacing of 0.5 pH units the exchange rate for the full-length ribozyme was increased to 40% (data not shown). As illustrated in Figure **3.2.2.1b**, reducing the pH-spacing more than doubled the transition rate in λ-space. We observed the most significant improvement in the transition rate of residue A17, which increased from 23ns^{-1} to 113 ns^{-1}. This is on par with the transition rate of 92 ns^{-1} observed in the GAAA tetraloop. The transition rate was also uniformly consistent across the 3 independent simulation runs, which ensures the robustness of the calculations. Qualitatively, the titration curves obtained across 3 independent runs also demonstrated better convergence for pH-REX simulations with smaller pH spacing. Finally, using this smaller pH spacing of 0.5, we reran pH-REX CPHMD$^{MS\lambda D}$ simulations on the full-length ribozyme. After an initial ~10 ns of equilibration, the calculated pK$_a$ values started to converge and results comparable to the GAAA tetraloop were achieved within 13 ns (**Table 3.2.2.1**), demonstrating that pH-REX CPHMD$^{MS\lambda D}$ simulation scales well to simulate pH-dependent properties of full-sized nucleic acid systems.

3.2.3 Conclusion: pH-REX Improves Sampling & Convergence in Explicit Solvent CPHMD$^{MS\lambda D}$ Simulations

In this chapter, we identified sampling challenges associated with modeling pH-dependent dynamics of RNA structures in explicit solvent. Consequently, we have enhanced the framework with pH-based replica exchange (pH-REX) sampling, which significantly improved sampling of titration and spatial coordinates, and the shuffling of conformations across pH space has the effect of decoupling interactions between titrating residues. This allows us to ameliorate some of the sampling issues related to orthogonal barriers that originate from coupled protonation equilibrium and conformational-dependent pK_a behavior, as illustrated in our example of the GAAA tetraloop motif, and this has the overall effect of improving accuracy from our initial results. The scalability of pH-REX sampling was also demonstrated by showing that similarly accurate pK_a values could be achieved when simulating full-sized nucleic acid systems, such as the lead-dependent ribozyme. Finally, we highlighted that pH-REX CPHMD$^{MS\lambda D}$ simulations can be used to identify the dominant conformation of nucleic acid structures in alternate pH environments or to provide structural characterization of pH-dependent transient states, making it a useful tool to provide accurate first-principles prediction, in terms of the structural and mechanistic insight into the study of pH-dependent properties of nucleic acids.

3.3 Explicit Solvent CPHMD of Proteins in CHARMM

Note: Chapter 3.3 was adapted from the following references.[114]

3.3.1 Optimization of Model Potential Parameters for 2-State Titrations

The existing generation of implicit solvent CPHMD in modeling the pH-dependent dynamics of proteins have met with considerable success. However, the generalized Born implicit solvent model is known to introduce systematic errors, such as underestimating desolvation energies of buried charge-charge interactions,[91] and causing structural compaction which may distort the overall protein structure.[96,103] In systems such as ion channels[104-106] and some transmembrane proteins,[107] microscopic interactions of discrete ions and water with the protein are important, the use of an explicit solvent representation of the solvent environment is desirable. Here, we extend the developed CPHMD$^{MS\lambda D}$ framework to model proteins. As with the previous implementation of CPHMD$^{MS\lambda D}$ for nucleic acids,[111] we used the free energy of deprotonation as the fixed biasing potential (F^{fixed}) in our simulation. The free energy of deprotonation was calculated for each isolated model compound embedded in explicit solvent using traditional λ-dynamics at zero ionic strength. In order to facilitate transitions between the two protonation states, we optimized the force constant (k_{bias}) on the variable biasing potential (F^{var}) that was applied to each model compound, and targeted to achieve a maximal value of the transition rate in λ-space (i.e., titration coordinate sampling), while maintaining a high fraction of physical ligands. The optimized parameters for the model potentials are reported in **Table 3.3.1.1**. Calculation of the sampling statistics (see **Table 3.3.1.2**) indicates that the fraction of physical states was maintained at ~70% and transitions in λ-space were ~50 transitions/ns. The sampling properties of our model amino acids are comparable to previous work performed on model nucleosides in explicit solvent.[111]

Next, we performed a 2-state titration simulation where only two titrating states (protonated and unprotonated) were simulated and tautomers for each protonation state were not explicitly modeled. The titration curves for our model compounds are illustrated in **Figure 3.3.1.1**. The calculated pK_a of aspartic and glutamic acid for a two-state titration (i.e., without proton tautomerism) were 4.1 and 4.3 pK_a respectively, which is within ±0.1 pK_a units from their experimental pK_a values of 4.0 and 4.4 respectively.[160] For the two-state titration of histidine, where either Nδ or Nε was titrated, the pK_a values obtained were 6.7 and 7.0 respectively,[161] which is identical to their experimental pK_a values. Finally, the calculated pK_a of lysine was 10.2, which is in close agreement with the experimental pK_a value of 10.4.[160] The excellent agreement between our model compounds calculated pK_a values and their experimental values indicate that the sampling of titration coordinates in our CPHMD$^{MS\lambda D}$ simulations was sufficient to yield accurate results.

Table 3.3.1.1: Parameters for the Model Potential for 2-state Titrations

Residue	**$\Delta G_{protonation}$ (kcal/mol)**	**F^{var} (kcal/mol)**	**Ref pK_a**
		k_{bias}	
Asp	43.71	34.00	4.00
Glu	46.00	34.25	4.40
His-δ	-3.58	26.00	7.00
His-ε	-12.26	26.00	6.60
Lys	-23.02	29.50	10.40

Table 3.3.1.2: Sampling characteristics of 2-state titration simulations performed at pH = pK_a.

Residue	**Fraction of Physical States**	**Transition (ns^{-1})**
Asp	0.74 ± 0.04	35 ± 5
Glu	0.75 ± 0.02	35 ± 5
His-δ	0.72 ± 0.03	60 ± 10
His-ε	0.71 ± 0.04	64 ± 14
Lys	0.76 ± 0.03	50 ± 8

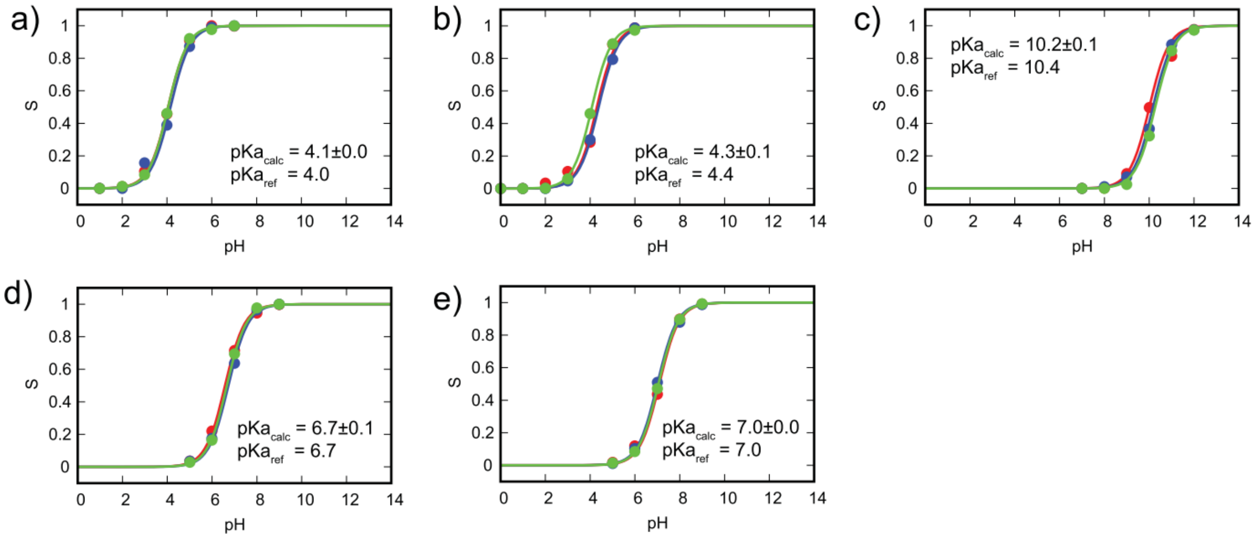


Figure 3.3.1.1: Titration curve of model compounds: (a) aspartic acid, (b) glutamic acid, (c) lysine, (d) (d) histidine-δ and (e) histidine-ε. Calculated pK$_a$ values of model compounds are in excellent agreement with experimental pK$_a$ values. Colors represent the results from the triplicate runs.

3.3.2 Optimization of Model Potential Parameters for 3-State Titrations

The original form of the F^{var} potential assumed the existence of only two states. When accounting for proton tautomerism and thus three states, the original form was not suitable because it frequently sampled an intermediate state of the two tautomers. This intermediate state is typically characterized by $\lambda_{\alpha,1} \approx 0$, $\lambda_{\alpha,2} \approx 0.5$ and $\lambda_{\alpha,3} \approx 0.5$, which corresponds to a half proton on both the Nδ and Nε protonation sites (using His as an example), and this represents an unphysical state whose sampling should be minimized. The existence of the intermediate state can be rationalized by considering that the free energy barrier for conversion between the two protonation states would be larger than the conversion between the two tautomers, as in the former process there is a change in the net charge of the system and a greater reorganization of the distribution of partial charges. The combined functional form of the original F^{var} potential that uses the same 0.8 cutoff in the definition of physical protonation states as expressed in eqn. 9, where $\lambda_{\alpha,1}$, $\lambda_{\alpha,2}$ and $\lambda_{\alpha,3}$ denote the alchemical scaling factors associated with each of the 3 states for some residue α, does not account for the uneven barrier height of the different alchemical reactions.

$$F_{\alpha}^{\mathrm{var}} = k_1\left(\lambda_{\alpha,1} - 0.8\right)^2 + k_1\left(\lambda_{\alpha,2} - 0.8\right)^2 + k_1\left(\lambda_{\alpha,3} - 0.8\right)^2 \tag{3.3.2.1}$$

To avoid the intermediate tautomeric states, we modified the existing F^{var} potential by including additional cross terms (k_2 expressions) to account for uneven barrier heights, and a final term (k_3 expressions) was added to ensure that the relative free energy of the end-states were not altered. The resulting functional form as outlined in **equation 3.3.2.2** results in a more versatile biasing potential that is suited to address the asymmetry of the potential energy surface associated with changes in both protonation and tautomeric states.

$$\begin{aligned} F_{\alpha}^{\mathrm{var}} = {} & k_1\left(\lambda_{\alpha,1} - 0.8\right)^2 + k_1\left(\lambda_{\alpha,2} - 0.8\right)^2 + k_1\left(\lambda_{\alpha,3} - 0.8\right)^2 \\ & + k_2\left(\lambda_{\alpha,1} - \lambda_{\alpha,2}\right)^2 k_2\left(\lambda_{\alpha,1} - \lambda_{\alpha,3}\right)^2 - k_3\left(\lambda_{\alpha,2}\right) - k_3\left(\lambda_{\alpha,3}\right) \end{aligned} \tag{3.3.2.2}$$

An iterative grid search strategy was used in testing various combinations of the force constants (k_1, k_2, k_3), and the optimal combination is reported in **Table 3.3.2.1**. As illustrated in **Figure 3.3.2.1**, which shows the time-evolution of λ, all 3 end states for the model compounds were well sampled. Calculation of the sampling statistics as summarized in **Table 3.3.2.2** indicates that the fraction of physical states was maintained above 70%, confirming that the modified F^{var} potential does not trap λ in an unphysical intermediate state. The transitions in λ-space were ~50 transitions/ns, which is comparable to the statistics obtained from 2-state titrations of the model compounds.

Table 3.3.2.1: Parameters for the Model Potential for 3-state Titrations

Residue	**$\Delta G_{\mathrm{protonation}}$ (kcal/mol)**	**F^{var} (kcal/mol)**			**Ref pK_a**
		k_1	**k_2**	**k_3**	
Asp-T	43.30	-16.5	18.5	-18.5	4.00
Glu-T	45.59	-16.0	18.5	-18.5	4.40
His-T	-3.58/-12.26	8.0	6.0	-6.0	6.45

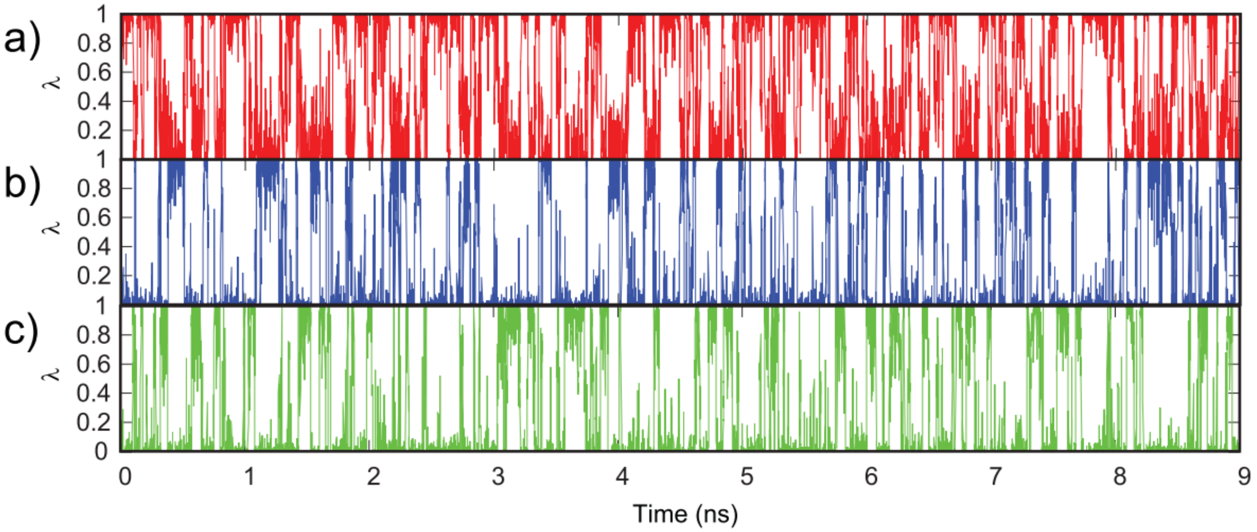


Figure 3.3.2.1: Titration coordinate transitions of aspartic acid at pH 4 for (a) unprotonated state, (b) protonated tautomer #1 and (c) protonated tautomer #2 shows that the physical end states are well sampled.

Table 3.3.2.2: Sampling characteristics of 3-state titration simulations performed at the pH closest to the model compound's pK_a value. Using pH-REX greatly accelerates sampling of titration coordinates with minimal loss in the fraction of physical states (FPS).

Residue	pH	Normal MD		pH-REX	
		FPS	Transition(ns^{-1})	FPS	Transition(ns^{-1})
Asp-T	4	0.78 ± 0.01	50 ± 1	0.78 ± 0.01	294 ± 16
Glu-T	4	0.76 ± 0.01	46 ± 7	0.75 ± 0.00	322 ± 21
His-T	7	0.81 ± 0.02	60 ± 6	0.81 ± 0.01	298 ± 12

While the sampling efficiency in λ-space of model compounds allows us to reproduce the pK_a values of the model compounds, the transition rate is nevertheless limited to ~50 transitions/ns. In our previous evaluation of explicit solvent CPHMD$^{MS\lambda D}$ simulations of larger nucleic acid structures, slower pK_a convergence was observed,[112] and it is likely that protein systems will encounter similar issues as well. The sampling of titration and spatial coordinates can be accelerated using a pH-REX sampling strategy.[113] Therefore, we have applied pH-REX sampling, and as illustrated in **Table 3.3.2.2**, it resulted in a 6-fold improvement in λ-space sampling of model compounds with effectively no loss in the fraction of physical states. As pH-REX sampling confers significant improvement over straightforward MD simulations and requires negligible overhead in terms of computational cost, the results presented in the subsequent sections are obtained from pH-REX CPHMD$^{MS\lambda D}$ simulations unless specified otherwise.

We performed a 3-state titration on the model compounds, where alchemical transformations across different protonation states and across different tautomers of the same protonation state were explicitly modeled. The tautomeric titrations of aspartic and glutamic yielded a pK_a of 4.4 and 4.8 respectively, which matches well with the macroscopic pK_a of 4.35 and 4.70 when the double degeneracy of the protonated states is taken into account.[90] However, since the experimental pK_a measured does not distinguish between the tautomeric forms, we recalibrated the fixed biasing potential in our CPHMD$^{MS\lambda D}$ simulations to reproduce the experimentally measured macroscopic pK_a values. This was achieved by reducing the biasing potential at pH=pK_a by $k_bT\ln(2)$ = 0.41 kcal/mol, which accounts for the degeneracy of the tautomeric protonated states. Our approach is different from that of Khandogin and co-workers,[90] where a post-correction factor of 0.3 pK_a units was applied to tautomeric residues. However, the net result in both approaches is the same, in the sense that the final pK_a value calculated accounts for tautomer degeneracy. The titration curves for our model compounds with proton tautomerism are illustrated in **Figure 3.3.2.2**. The calculated pK_a of aspartic and glutamic acid was 4.17 and 4.37 respectively, which is good agreement with experimental pK_a values. For histidine tautomeric titration, no re-calibration was performed because the pK_a measured by experiments were microscopic pK_a associated with the titration at the Nε and Nδ sites, and the fixed biasing potential applied to each tautomer was identical to those used in the 2-state titration setup. Our calculated pK_a for the histidine tautomer was 6.45, which is identical to the expected macroscopic pK_a value of 6.45.[90]

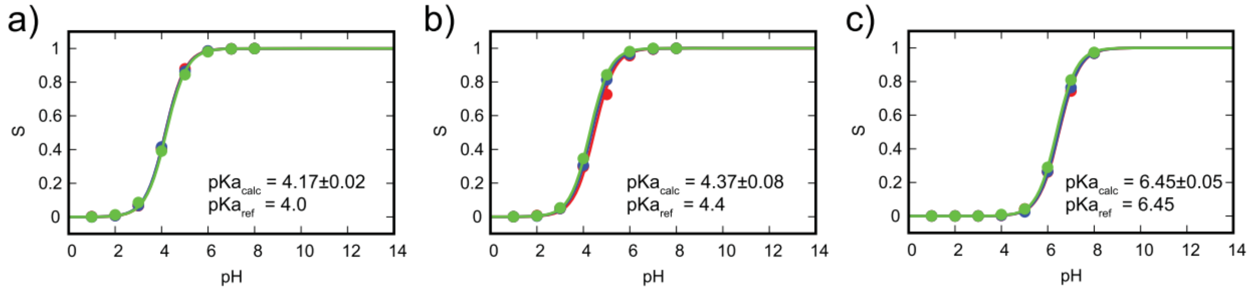


Figure 3.3.2.2: Titration curve of model compounds with proton tautomerism: (a) aspartic acid, (b) glutamic acid and (c) histidine. Calculated pK_a values of model compounds show excellent agreement with experimental pK_a values. Colors represent the results from the triplicate runs.

3.3.3. Modeling Interactions between Adjacent Titrating Residues

Further validation of the CPHMD$^{MS\lambda D}$ framework was performed on model dipeptide sequences Asp-Asp, Glu-Glu and Lys-Lys at zero ionic strength, where both residues were titrated simultaneously. The calculated pK_a values are summarized in **Table 3.3.3.1**. For the aspartic acid dipeptide, we observed that the pK_a values were 3.1 and 4.6, with the N-terminus Asp having a consistently lower pK_a in all 3 simulations runs, suggesting that the two Asp residues are in a different electrostatic environment. An analysis of the hydrogen bonding contacts that each Asp side chain forms with the backbone of the dipeptide (data not shown) indicated that the N-terminus Asp had 3 hydrogen bond donors within a ~5 Å radius, compared to the C-terminus Asp that had only 2 hydrogen bond donors. Thus, the increased presence of hydrogen bond donors around the N-terminus Asp facilitated the stabilization of its charged unprotonated state, explaining the decrease of its calculated pK_a value. By contrast, the calculated pK_a values for the glutamic acid dipeptide was 4.3 for both residues with no apparent pK_a shift. Similarly, the pK_a values for the lysine dipeptide was ~10.4 for both residues. The identical pK_a for both N- and C-terminus residues of both Glu-Glu and Lys-Lys dipeptides suggest that the electrostatic environment around each residue is similar. This is supported by the observation that no hydrogen bonding capable backbone atom was present in a ~5 Å proximity from the titrating functional group, and so the backbone interactions that were responsible for creating an asymmetric environment in Asp-Asp is significantly reduced in both Glu-Glu and Lys-Lys dipeptides.

Table 3.3.3.1: Calculated pK_a of various model dipeptide sequences. Values reported in the top table were calculated using **equation 2.2.3.2** (identity of residue was pre-assigned), and those reported in the bottom table were calculated using **equation 2.2.4.1** (identity of residue was not pre-assigned).

Residue Identity Pre-Assigned					
Residue	**Ref pK_a (of amino acid)**	**Site1**		**Site2**	
		pK_a	**n**	**pK_a**	**n**
Asp-Asp	4.0	3.1 ± 0.2	0.7 ± 0.1	4.6 ± 0.1	0.7 ± 0.0
Glu-Glu	4.4	4.3 ± 0.1	0.7 ± 0.0	4.3 ± 0.1	0.7 ± 0.0
Lys-Lys	10.4	10.3 ± 0.0	0.7 ± 0.0	10.5 ± 0.1	0.7 ± 0.0
Residue Identity Not Pre-Assigned					
Residue	**Ref pK_a (of amino acid)**	**Site1[a]**		**Site2[a]**	
		pK_a	**n**	**pK_a**	**n**
Glu-Glu	4.4	3.6 ± 0.0	-	5.0 ± 0.0	-
Lys-Lys	10.4	9.8 ± 0.1	-	11.0 ± 0.1	-

[a] Site1 and Site2 pK_a values are defined as the residue that produces the lower and higher "instantaneous" pK_a value. When averaged across the entire trajectory, they would correspond to the two macroscopic pK_a values recorded by experiments.

The calculated Hill coefficients of 0.7 suggest that anti-cooperative coupling is the dominant mode of interaction between the two adjacent titrating residues of these dipeptide systems. Prior work by Bashford and Karplus has demonstrated that when two residues titrate in the same pH region and have the same intrinsic (microscopic) pK_a, such as the Glu-Glu and Lys-Lys dipeptides in this analysis, the magnitude of their coupled interaction can raise/lower the apparent (macroscopic) pK_a of the system.[140] The existing HH-equation (i.e. **equation 2.2.3.2**) which we fitted our data to calculate a pK_a value is a rearranged form of the equation first proposed by Tanford and Roxby.[162] When there is no coupling with other titrating residues (i.e. $n = 1$), **equation 2.2.3.2** reduces to a form that can be derived from a mean-field approximation.[140] When there is coupling with other titrating residues (i.e. $n \neq 1$), the convention is to add the Hill coefficients to describe the anti-cooperative proton binding behavior. However, prior work by

Onufriev *et. al.* in their derivation of the decoupled site representation (DSR) framework has shown that this approach may not give the best fit to experimental macroscopic pK_a values.[138] Consequently, it is not unexpected that our analysis was unable to obtain the macroscopic pK_a of the Glu-Glu and Lys-Lys dipeptides, where one would expect to see two distinct pK_a values. If one wishes to elucidate the coupled pK_a behavior for these two dipeptides, the pK_a values can be recalculated by fitting it to a modified version of the HH-equation, which can be derived from the DSR approach.[138] In this revised fitting method using **equation 2.2.4.1**, where we analyzed the net proton uptake without pre-assigning the identity of each residue, the apparent pK_a values calculated cannot be assigned to a specific titrating site (i.e. the calculated pK_a values are not the microscopic pK_a of specific residues). Using this approach, two clear and distinct pK_a values emerge for Glu-Glu (3.6, 5.0) and Lys-Lys (9.8, 11.0), which is consistent with the perturbation of one protonated residue on the other.

3.3.4 pK_a calculations of Model Proteins using CPHMD$^{MS\lambda D}$ simulations

The HEWL protein is a well-studied protein system that contains the 3 most common titrating residues (Asp, Glu, His) with site-specific pK_a values for each residue that have been measured in a number of experimental studies.[163-168] It is perhaps the closest thing to a "universal benchmark" system that has been evaluated by numerous CPHMD implementations over the years.[72,77,78,80,91,103,169] To the best of our knowledge, all existing "pure" explicit solvent CPHMD simulations reported in the literature have only been demonstrated on small peptide compounds[109] and simple organic molecules.[170] We performed a 20 ns pH-REX CPHMD$^{MS\lambda D}$ simulation of HEWL, which is the first example of explicit solvent CPHMD simulation on a full protein to be reported.

pK$_a$ calculations over 5 ns interval segments of our pH-REX CPHMD$^{MS\lambda D}$ trajectory show that good convergence is achieved within 20 ns. The difference in pK$_a$ values across our triplicate runs is small, typically between 0.2 to 0.3 pK$_a$ units, demonstrating that our results are robust and reproducible. The accuracy of our calculated pK$_a$ values are then compared to experimental measurements from consensus NMR titrations.[168] As summarized in **Table 3.3.4.1**, the calculated pK$_a$ values are in good agreement with experiment, with a root-mean-square-error (RMSE) of 0.85 pK$_a$ units and an average unsigned error (AUE) of 0.68 pK$_a$ units. Nielson and co-workers previously estimated that experimental pK$_a$ values reported in the literature on average may vary by 0.5 pK$_a$ units depending on the experimental method and/or protocol used to make the measurements.[168] This suggests that the accuracy of our pH-REX CPHMD$^{MS\lambda D}$ simulations are approaching the uncertainty of experimental pK$_a$ values.

Table 3.3.4.1: pK$_a$ values of HEWL calculated using implicit and hybrid solvent pH-REX CPHMD simulations as reported by Wallace and Shen, compared to pK$_a$ values calculated using explicit solvent pH-REX CPHMD$^{MS\lambda D}$ simulations in this work. Calculated pK$_a$ values with error greater than 1.0 pK$_a$ unit relative to experimental values based on consensus NMR titrations are identified in red.

Residue	**Exp pK$_a$**	**Implicit CPHMD**		**Hybrid CPHMD**		**Explicit CPHMD$^{MS\lambda D}$**	
		pK$_a$	**Error**	**pK$_a$**	**Error**	**pK$_a$**	**Error**
GLU-7	2.6 ± 0.2	2.6 ± 0.1	0.0	2.7 ± 0.0	0.1	2.7 ± 0.1	0.1
HIS-15	5.5 ± 0.2	5.3 ± 0.5	-0.2	6.6 ± 0.1	1.1	6.0 ± 0.2	0.5
ASP-18	2.8 ± 0.3	2.9 ± 0.0	0.1	3.1 ± 0.1	0.3	2.1 ± 0.2	-0.7
GLU-35	6.1 ± 0.4	4.4 ± 0.2	-1.8	7.2 ± 0.2	1.1	7.0 ± 0.3	0.9
ASP-48	1.4 ± 0.2	2.8 ± 0.2	1.4	1.6 ± 0.5	0.2	1.3 ± 0.0	-0.1
ASP-52	3.6 ± 0.3	4.6 ± 0.0	1.0	2.9 ± 0.1	-0.7	4.5 ± 0.3	0.9
ASP-66	1.2 ± 0.2	1.2 ± 0.4	-0.1	1.5 ± 0.6	0.3	1.5 ± 0.1	0.3
ASP-87	2.2 ± 0.1	2.0 ± 0.1	-0.2	1.5 ± 0.4	-0.7	1.3 ± 0.0	-0.9
ASP-101	4.5 ± 0.1	3.3 ± 0.3	-1.2	3.0 ± 0.1	-1.5	5.1 ± 0.5	0.6
ASP-119	3.5 ± 0.3	2.5 ± 0.1	-1.1	2.9 ± 0.1	-0.7	1.6 ± 0.0	-1.9
RMSE		**0.94**		**0.80**		**0.84**	
AUE		**0.70**		**0.66**		**0.68**	

Next, we identified the residues that had errors in their calculated pK$_a$ values, which we defined as having more than 1.0 pK$_a$ unit difference between calculated and experimental values.

Asp-119 was underpredicted by -1.9 pK_a units, which suggests that the unprotonated state is overstabilized in our simulations. Analysis of its microenvironment indicates that persistent hydrogen bond interactions between the carboxylic oxygens of Asp-119 and the amide backbone hydrogen of Gln-121 and Ala-122 were present even in a low pH environment, which accounts for the extra stabilization of the unprotonated state of Asp-119. Similar underprediction of Asp pK_a values has been documented in other CPHMD work, where salt bridge interactions were responsible.[78] When non-salt-bridge configurations were sampled, it resulted in more accurate pK_a results.[78] This suggests that the apparent error in the Asp-119 pK_a value could be a sampling issue, and more extensive sampling or more aggressive sampling methods may be required when dealing with residues that are "locked" to their initial conformation by strong interactions like hydrogen bonds or salt bridges.

3.3.5 Comparison to Implicit Solvent CPHMD Simulations

We compared the performance of explicit solvent pH-REX CPHMD$^{MS\lambda D}$ simulations to CPHMD models implemented in other solvation models. A number of CPHMD variations have been implemented in AMBER[77] and GROMACS.[72] However, they will not be included our analysis as deconvoluting the effects originating from force field differences to those arising from solvation model differences is not straightforward. Instead, we will focus our analysis on CPHMD variations implemented in CHARMM. The original CPHMD in CHARMM was implemented with a GB implicit solvent model,[89] and we have used the HEWL pK_a values reported by Wallace and Shen for comparison.[103] Since that work was reported using a pH-REX sampling strategy, we have also eliminated the effects of using different sampling strategies. At the time of writing, there is no "pure" explicit solvent CPHMD based on the CHARMM force field that has been tested on the HEWL protein. However, a close comparison can be made with

Shen's hybrid solvent CPHMD model.[103] The key methodological difference between explicit and hybrid solvent models is that the evaluation of free energies of deprotonation and the forces on the fictitious λ particles that govern the titration coordinates are calculated using a GB implicit solvent model in Shen's hybrid solvent CPHMD model, whereas in our explicit solvent CPHMD$^{MS\lambda D}$ model there is no use of the GB implicit solvent model in any part of the calculation. Unfortunately, the use of such hybrid sampling means there is no clear Hamiltonian for this system and correspondence to results from any specific statistical mechanical derivation cannot be demonstrated. Lastly, the sampling of titration coordinates in implicit solvent is typically ~2000 transitions/ns,[80] which is an order of magnitude higher than those obtained in our explicit solvent simulations. Therefore, to compensate for the differential sampling speed associated with different solvent models, we compared the results of our 20 ns pH-REX CPHMD$^{MS\lambda D}$ trajectories to the previously reported 2 ns pH-REX trajectories that uses the implicit and hybrid CPHMD model.

As summarized in **Table 3.3.4.1**, in terms of overall pK_a predictive performance, our explicit solvent CPHMD$^{MS\lambda D}$ results had a RMSE error of 0.84 pK_a units. This is an improvement from the results obtained using implicit solvent CPHMD (RMSE = 0.94), and our model performance is close to that of the hybrid solvent CPHMD (RMSE = 0.80). A similar trend was also noted using alternative error metrics, such as the average unsigned error (AUE). We then identified the number of residues that had errors of more than 1.0 pK_a unit relative to experimental values. Our explicit solvent CPHMD$^{MS\lambda D}$ model had only 1 such residue (i.e., Asp-119) compared to the implicit and hybrid solvent CPHMD models which had 5 and 3 residues respectively. Notable improvements in moving from a hybrid solvent to a "pure" explicit solvent model can be observed in His-15, where the overestimation of its pK_a value is reduced from 1.1

to 0.5 pK_a units. Similarly, the hybrid solvent CPHMD model incorrectly predicted the direction of pK_a shift for residue Asp-101, whereas the explicit solvent CPHMD$^{MS\lambda D}$ model not only predicted the right direction of pK_a shift, but the magnitude of error was also smaller (-1.5 vs +0.6). Our findings suggest that when corrected for differences in titrating coordinates sampling, the explicit solvent CPHMD$^{MS\lambda D}$ model produces more accurate pK_a predictions than the original implicit solvent CPHMD.

3.3.6. Generalizability to Other Proteins

Lastly, to demonstrate that the pK_a calculations obtained from the CPHMD$^{MS\lambda D}$ framework for proteins is not specific to HEWL protein, we performed pK_a calculations on two additional proteins, the BBL and NTL9 protein. Given that we have only investigated a single His residue in a protein environment, for BBL we only titrated the two His residues. NTL9 has no His residues, and the Glu and Asp residues that have experimental pK_a measurements were titrated.

Table 3.3.6.1: pK_a values of BBL and NTL9 calculated using explicit solvent pH-REX CPHMD$^{MS\lambda D}$ simulations in this work. Calculated pK_a values with error greater than 1.0 pK_a unit relative to experimental values[109,110] based are identified in red.

Residue	**Explicit CPHMD$^{MS\lambda D}$**		
	Exp pK_a	**pK_a**	**Error**
		BBL	
HIS-142	6.5	6.6 ± 0.1	0.1
HIS-166	5.4	4.8 ± 0.0	-0.6
		NTL9	
ASP-8	3.0	1.5 ± 0.1	-1.5
GLU-17	3.6	4.0 ± 0.5	0.4
ASP-23	3.1	3.7 ± 0.2	0.6
GLU-38	4.0	3.9 ± 0.2	-0.1
GLU-48	4.2	3.4 ± 0.3	-0.8
GLU-54	4.2	3.6 ± 0.2	-0.6
RMSE			**0.72**
AUE			**0.59**

As summarized in **Table 3.3.6.1** the calculated pK_a values have are reasonably accurate (RMSE = 0.72, AUE = 0.59).[171,172] From the experimental data, most of the residues titrate close to the pK_a of their reference compounds, but two residues had more than a 1.0 pH unit shift. For His-166 of BBL, the residue is buried and its experimental pK_a is 5.4. For Asp-8 of NTL9, its experimental pK_a of 3.0 can be traced to the salt bridge interactions it forms with the amide backbone of adjacent residues. Our calculated pK_a values demonstrate a similar downward shift, although in both cases the extent of the shift tends to be overestimated. We suggest that this overestimation may be due to the lack of sampling stemming from the shorter 3 to 5 ns simulations performed for these systems. In other proteins like staphylococcal nuclease, residues with shifted pK_a values are known to undergo local conformational changes,[173,174] and sampling these states will be required to improve the accuracy of pK_a calculations. Together with our observations for Asp-119 in HEWL, our work suggests that while short pH-REX CPHMD$^{MS\lambda D}$ simulations are capable of reproducing experimental pK_a values of most protein residues, accurate reproduction of highly shifted pK_a values (e.g., buried charged residues) or those involving salt-bridge or similarly strong interactions remains a challenge that may be better addressed with more aggressive conformational sampling techniques.

3.3.7. Updated Model Potential for CHARMM36

With the development of the CHARMM36 force field for proteins,[130,131] which corrects the balance between α-helix and β-sheet structures, a substantial proportion of the dihedral parameters were modified from the previous CHARMM22 force field. In CPHMD, the calculated free energy of deprotonation not only comprises of the energy differences arising from the differences in the electrostatics and van der Waals terms, but those of the internal bonded terms as well. Therefore, we report the following updated model potentials for use with the

CHARMM36 force field for proteins (**Table 3.3.7.1**). In addition, model potential parameters were constructed for titrating Tyrosine, which in recent studies has also been implicated in pH-dependent protein activity.[175]

Table 3.3.7.1: CPHMD parameters for CHARMM36 compatible model compounds

	$\Delta G_{protonation}$ (kcal/mol)	**Fvar (kcal/mol)**			**Ref pK$_a$**
		k_1	**k_2**	**k_3**	
Asp-T	51.28	-19.25	21.25	-21.25	4.00
Glu-T	53.81	-19.25	21.50	-21.50	4.40
His-T	-2.62/-13.67	8.75	6.75	-6.75	6.45
Tyr	102.57	35.00	-	-	9.60
Lys	-30.63	35.00	-	-	10.40

3.3.8 Conclusion: Implemented an Explicit Solvent CPHMD$^{MS\lambda D}$ Framework for Proteins

In this chapter, we have extended the existing explicit solvent CPHMD$^{MS\lambda D}$ framework to simulate the pH-dependent properties of proteins, by developing the appropriate model potential parameters for amino acids model compounds. In the CPHMD$^{MS\lambda D}$ framework, we performed seamless alchemical transitions between protonation and tautomeric states using multi-site λ-dynamics, and designed a novel biasing potential to ensure that only the physical end-states are predominantly sampled. We also determined the proper treatment for dealing with coupled titrating systems where the identity of various residues cannot be pre-determined, which underscores the distinction between microscopic vs macroscopic pK$_a$ measurements. In addition, we have demonstrated the first examples of a “pure” explicit solvent CPHMD simulations to simulate realistic pH-dependent properties of a number of model full-sized protein systems, including HEWL, BBL and NTL9. Our pK$_a$ calculations for HEWL protein are in excellent agreement with experimental values, with a RMSE of 0.84 pK$_a$ units, and this is close to the uncertainty of 0.50 pK$_a$ units associated with experimental measurements. With the development of explicit solvent CPHMD$^{MS\lambda D}$ for proteins, it will finally allow us to address questions related to pH-dependent properties of membrane proteins and ion channels, where discrete representation of ions and water is important.

3.4 Explicit Solvent CPHMD of Nucleic Acids in AMBER

Note: Chapter 3.4 was adapted from the following references.[115,116] Chapter 3.4.2 to 3.4.4 contains considerable contributions from Kamali Sripathi, who was responsible for parameterizing the alternative protonation states in the AMBER force field.

3.4.1. Performance of CHARMM 36 Nucleic Acid Force Field on RNA Structures

Progressive work in using the explicit solvent CPHMD framework to model increasingly complex RNA systems led to the observation that the current CHARMM36 nucleic acid force field may not be sufficiently optimized for modeling all but ideal DNA/RNA helices. In the context of pH-dependent RNA activity, structural features of interest include bulges,[35] triplexes[21,26,176-178] and pseudoknot structure[21,26,176-178] that contains protonated residues, and using the CHARMM36 force field to model these type of structural motifs have not been extensively validated in the literature. Preliminary data suggests that over a longer timescale, typically on the order of >100 ns, the protonated form of these RNA structures deviate significantly from their native crystallographic or NMR structures. For example, protonated $A{\bullet}C^+$ base pairs that are known to be stable lose their base-paired conformation within 100 ns (data not shown). Extensive validation efforts have also indicated that the partial charge parameterization scheme is not the main determining factor for the inability to maintain correct geometries of protonated base pairs, as the interaction energies of protonated $A{\bullet}C^+$ base pairs, a proxy of how "strong" the base pairing interaction should be, did not have any discernable effect on the ability of the force field to maintain a correct protonated base-pair geometry (data not shown).

MacKerell and co-workers have previously reported that sampling the 150° to 250° region of the 2'-hydroxyl dihedral phase space is correlated to RNA conformational heterogeneity, specifically in promoting the formation of non-canonical conformations.[179] The previous CHARMM27 nucleic force field was known to oversample this region of phase space, and it was the primary reason as to why simulated structures could not maintain their structural integrity. In the context of our studies, we observed that the loss of native structure correlates to the 2'-hydroxyl dihedral sampling in the 150° to 250° region, which led to the hypothesis that the dihedral parameters were the primary cause for the inability to maintain correct geometries of protonated RNA structures. Even though the CHARMM36 force field was developed to minimize this occurrence of non-canonical tertiary structure sampling, our data suggests that it may not be applicable to nucleic acid structures that possess more complex tertiary structure, beyond the typical A-form helix for RNA. In the development of the CHARMM36 force field, several candidates, CHARMM27a through CHARMM27e were reported and tested by MacKerell and co-workers.[179] Ultimately, the CHARMM27d force field was selected as the CHARMM36 force field. However, we noted that both CHARMM27b and CHARMM27d force field demonstrated almost equivalent performance in the 3 benchmark studies performed – water probability overlap, J-coupling and free energy calculations. However, the CHARMM27b force field stabilizes the 50° to 100° region of the 2'-hydroxyl dihedral phase space more strongly than the CHARMM27d (CHARMM36) force field, which is the region that promotes canonical tertiary structure. Preliminary tests demonstrated that using the CHARMM27b force field did produce a notable improvement in the stability of protonated RNA structures in its ability to maintain protonated A•C^+ base pairs for a longer time (data not shown), which suggests that the 2'-hydroxyl dihedral parameter will need to be further optimized to model protonated RNA structures accurately. As a temporary measure, we recommend that the CHARMM27b parameters should be used instead of the official CHARMM36 parameters when simulating nucleic acid structures that have segments that are not A-form helical.

As the focus of this dissertation is on the development and application of explicit solvent CPHMD, as opposed to nucleic acid force field development and validation, a decision was made to utilize the AMBER parmbsc0 nucleic acid force field with χ_{OL3}-correction,[180] which has received more developmental effort, and has been successfully used to model a number of complex RNA structures.[181,182]

3.4.2. New AMBER Parameters for Protonated Nucleic Acids

Here, we summarize different parameterization schemes tested for the purpose of parameterizing alternative protonation states of nucleobases and related compounds for the AMBER nucleic acid force field. In the standard AMBER parameterization protocol, the partial charge distribution is obtained from the RESP charges from QM calculations using HF/6-31G*.[183] As protonated bases are more subject to polarizable effects than neutral bases, we investigated if using a higher level of theory, more comprehensive basis sets that add diffuse functions, and/or performing the QM calculations under different dielectric environments would affect the accuracy of simulations of protonated RNA structures.

To validate simulation accuracy, we used the protonated A25•C6$^+$ base pair in the lead-dependent ribozyme as a simple model system. As an initial screen, we first calculated the interaction energy of protonated A25•C6$^+$ base pair. The results as summarized in **Table 3.4.2.1**, indicate that the interaction energies across solvent environments, while controlling for the same level of theory and basis set, were consistent with one another, and varied by only 2-3 kcal/mol. In light of the insensitivity of interaction energies to the solvent environment at which the QM calculations were performed, we used the gas phase calculations for the next validation stage, as this is most consistent with the standard AMBER parameterization protocol.[183] Next, we tested the parameters obtained from 18 permutations of level of theory and basis sets in their ability to reproduce the experimental pK_a of 6.5. All pK_a values calculated were within a standard deviation of ±1 pK_a unit of the mean value (**Table 3.4.2.2**).

Table 3.4.2.1: Interaction energies of the A25•C6$^+$ base pair in lead-dependent ribozyme, using parameters from QM calculations of varying level of theory, basis set and solvent environment.

Theory	Basis Set	Gas	Ether	Water
HF	6-31G*	-39.8630	-41.9912	-43.2652
	6-31+G*	-40.6766	-42.7231	-43.8629
	cc-pVDZ	-38.9907	-41.6234	-43.3622
	aug-cc-pVDZ	-40.4656	-41.9959	-43.2998
	cc-pVTZ	-38.9157	-41.7731	-41.9168
	aug-cc-pVTZ	-38.8047	-41.7977	-42.5464
B3LYP	6-31G*	-38.4736	-40.7321	-41.8712
	6-31+G*	-38.7940	-41.3145	-42.5700
	cc-pVDZ	-38.1606	-39.7476	-40.8145
	aug-cc-pVDZ	-38.2135	-40.4519	-42.1383
	cc-pVTZ	-37.7925	-40.2338	-41.234
	aug-cc-pVTZ	-37.7057	-40.4711	-42.0726
MP2	6-31G*	-39.8630	-41.9912	-43.2652
	6-31+G*	-40.6766	-42.7231	-43.8629
	cc-pVDZ	-40.1295	-41.1844	-42.8962
	aug-cc-pVDZ	-40.4656	-41.9959	-43.2998
	cc-pVTZ	-38.9157	-41.7731	-41.9168
	aug-cc-pVTZ	-38.8047	-41.7977	-42.5464

Table 3.4.2.2: pK_a calculations of the A25•C6$^+$ base pair in lead-dependent ribozyme, using parameters from QM calculations of varying level of theory and basis set.

Theory	Basis Set	Calculated pKa
HF	6-31G*	6.7
	6-31+G*	6.9
	cc-pVDZ	7.5
	aug-cc-pVTZ	6.5
	cc-pVTZ	6.2
	aug-cc-pVTZ	6.4
B3LYP	6-31G*	6.8
	6-31+G*	7.0
	cc-pVDZ	6.1
	aug-cc-pVDZ	6.6

	cc-pVTZ	6.7
	aug-cc-pVTZ	6.8
MP2	6-31G*	6.8
	6-31+G*	6.9
	cc-pVDZ	6.8
	aug-cc-pVDZ	7.2
	cc-pVTZ	6.6
	aug-cc-pVTZ	7.4

Based on our results, it is apparent that changing the level of theory, basis sets and dielectric environment in the QM calculations used to derive the parameters for alternative protonation state residues does not have a discernable effect on the overall accuracy of reproducing experimental pH-dependent observables. To maintain consistency with the standard parameterization AMBER protocol, we utilized the parameters obtained at the Hartree-Fock (HF) level of theory and the 6-31G* basis set, calculated in gas phase. Specifically, to determine the partial charge parameters for alternative protonation states of nucleobases, we used methylated nucleobases as the model compound. The partial charges used directly corresponded to the RESP calculated charges without modification for all atoms, with the exception of atom N9 that collected the residual charge between the sum of RESP charges and the total expected charge of the base fragment, which is necessary to maintain an integer charge for the entire fragment. For parameterizing alternative protonation states of the 2'OH group of ribose sugars, we constrained the partial charges of the upper sugar fragment obtained from the AMBER force field in its standard protonation state, and rescaled the RESP charges of the lower sugar by normalizing it to the total expected charge of the fragment. For the model compound ribose sugar used calibrating our CPHMD$^{MS\lambda D}$ simulations, we constrained the partial charges of the lower sugar and atom O4' and rescaled the RESP charges of the upper sugar by normalizing it to the total expected charge of the fragment. The final set of partial charge parameters developed are summarized in **Table 3.4.2.3.**

Table 3.4.2.3: Partial charge distribution of AMBER-compatible parameters of alternative protonation states.

Residue	Atom Name	Atom Type	Charge	Residue	Atom Name	Atom Type	Charge
Ade (RNA)	N9	NS	0.0392	**Ade (DNA)**	N9	NS	0.0580
	C8	C2	0.1328		C8	C2	0.1328
	H8	H5	0.2014		H8	H5	0.2014
	N7	NB	-0.5216		N7	NB	-0.5216
	C5	CB	0.2367		C5	CB	0.2367
	C6	CA	0.2758		C6	CA	0.2758
	N6	N2	-0.7283		N6	N2	-0.7283
	H61	H	0.4298		H61	H	0.4298
	H62	H	0.4298		H62	H	0.4298
	N1	NC	-0.2473		N1	NC	-0.2473
	H1	H	0.3515		H1	H	0.3515
	C2	CQ	0.1901		C2	CQ	0.1901
	H2	H5	0.1921		H2	H5	0.1921
	N3	NC	-0.4582		N3	NC	-0.4582
	C4	CB	0.3521		C4	CB	0.3521
Cyt (RNA)	N1	NS	0.1418	**Cyt (DNA)**	N1	NS	0.1722
	C6	C1	-0.0374		C6	C1	-0.0374
	H6	H4	0.2414		H6	H4	0.2414
	C5	CM	-0.3838		C5	CM	-0.3838
	H5	HA	0.2190		H5	HA	0.2190
	C4	CA	0.6096		C4	CA	0.6096
	N4	N2	-0.9084		N4	N2	-0.9084
	H41	H	0.4753		H41	H	0.4753
	H42	H	0.4753		H42	H	0.4753
	N3	NC	-0.3288		N3	NC	-0.3288
	H3	H	0.3388		H3	H	0.3388
	C2	C	0.5419		C2	C	0.5419
	O2	O	-0.4782		O2	O	-0.4782
Gua (RNA)	N9	NS	-0.0546	**Gua (DNA)**	N9	NS	-0.0397
	C8	CK	0.0358		C8	CK	0.0358
	H8	H5	0.1301		H8	H5	0.1301
	N7	NB	-0.5707		N7	NB	-0.5707
	C5	CB	0.0197		C5	CB	0.0197
	C6	C	0.7437		C6	C	0.7437
	O6	O	-0.6998		O6	O	-0.6998

	N1	NA	-0.8598		N1	NA	-0.8598
	C2	CA	0.8907		C2	CA	0.8907
	N2	N2	-0.9641		N2	N2	-0.9641
	H21	H	0.3611		H21	H	0.3611
	H22	H	0.3611		H22	H	0.3611
	N3	NC	-0.8007		N3	NC	-0.8007
	C4	CB	0.3038		C4	CB	0.3038
Ura (RNA)	N1	NS	-0.1875	**Thy (DNA)**	N1	NS	-0.1492
	C6	CM	-0.1156		C6	CM	-0.3202
	H6	H4	0.1593		H6	H4	0.1983
	C5	CM	-0.5167		C5	CM	-0.0638
	H5	HA	0.1528		C7	CT	-0.2203
	C4	C	0.9652		H71	HC	0.0512
	O4	O	-0.7610		H72	HC	0.0512
	N3	NC	-0.9497		H73	HC	0.0512
	C2	C	0.8631		C4	C	0.8263
	O2	O	-0.7437		O4	O	-0.7225
2'OH (RNA)	C2'	CT	0.3033		N3	NC	-0.9187
	H2'	H1	-0.1541		C2	C	0.8210
	O2'	OX	-1.1803		O2	O	-0.7313
Phos (Taut #1)	P	P	1.2181	**Phos (Taut #2)**	P	P	1.2181
	O1P	OH	-0.5878		O1P	O2	-0.5769
	O2P	O2	-0.5769		O2P	OH	-0.5878
	H1P	HO	0.5603		H2P	HO	0.5603

3.4.3 An Expanded Repertoire of Titratable Groups for Nucleic Acids

In the explicit solvent CPHMD$^{MS\lambda D}$ framework of nucleic acids using the AMBER force field, we expanded the titratable groups to encompass all major titration sites. As illustrated in **Figure 3.4.3.1**, this includes titration of all 5 bases – Ade, Cyt, Gua, Ura, Thy at the N1/N3 protonation site, backphone phosphate and its protonation site on the non-bridging oxygen and the 2'OH protonation site in ribose sugar. While Ade and Cyt are the most commonly observed protonation sites with shifted pKa, the expanded titrating functionality developed will allow us to address a larger range of systems, including systems where Gua titration may be important,[40,42,184] as the well as the backbone phosphate and 2'OH group that are implicated in the catalytic mechanism of a number of ribozyme systems.

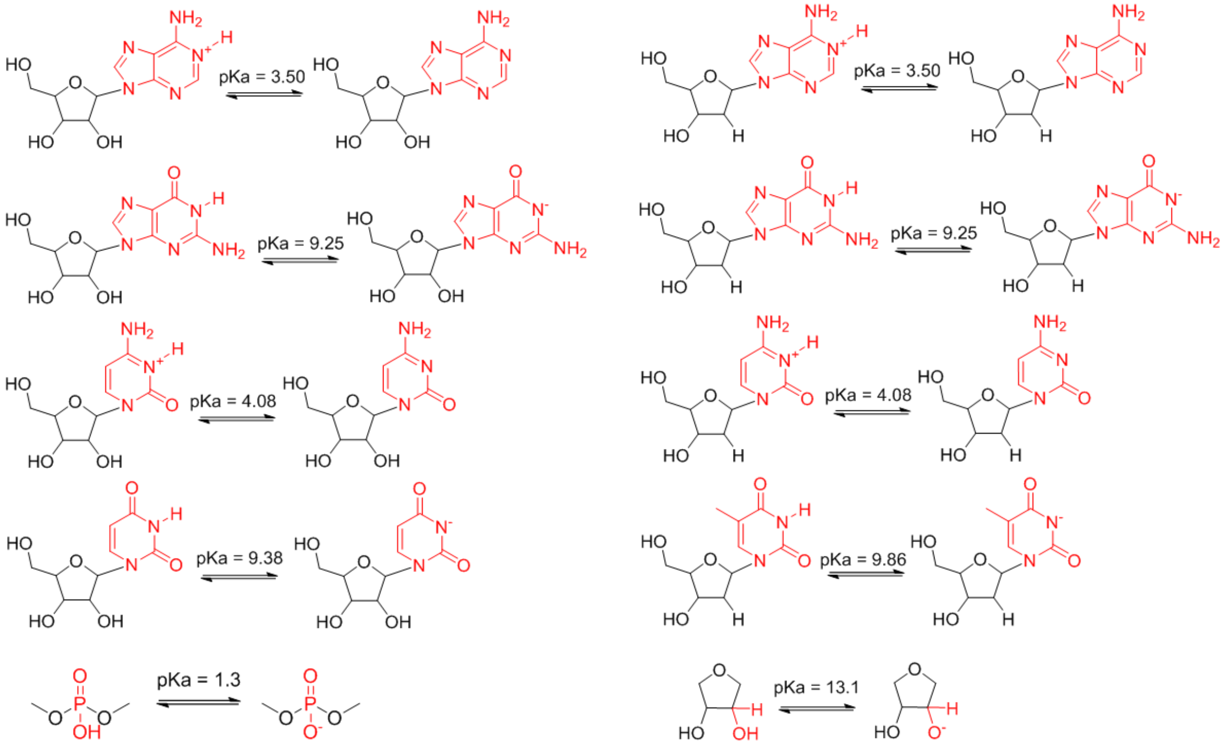


Figure 3.4.3.1: List of titratable groups and residues in the expanded set of nucleic acid CPHMD

As with the methodology outlined in our previous work,[111] we calculated the model potential parameters for each residue, which is summarized in **Table 3.4.3.1**. In addition, for the backbone phosphate group, because of the tautomeric states for the protonated non-bridging oxygen, a modified functional form of the fixed biasing potential was applied to allow tautomeric titrations (see **Table 3.4.3.1**), and this approach is adopted from our previous work of modeling tautomeric states in protein side chains.[114]

Table 3.4.3.1: Parameters for the Model Potential for 2-state Titrations

Residue	**$\Delta G_{protonation}$ (kcal/mol)**	**Fvar (kcal/mol)**			**Ref pK_a**
		k_1	**k_2**	**k_3**	
Ade	73.1	24.00	-	-	3.50
Cyt	52.6	26.25	-	-	4.08
Gua	80.4	27.50	-	-	9.25
Ura	87.6	28.25	-	-	9.38
2’OH	-102.5	54.50	-	-	13.10
Phos	19.6	-26.4 /	24.8	-24.8	1.30

3.4.4 pK_a calculations of Model RNA systems

We further tested the applicability the newly developed parameters by reproducing a broad array of pK_a values. While site-specific pK_a measurements of protonated residues of nucleic acids are sparse in the literature, we managed to select a group of 5 RNA and DNA structures that have diverse local microenvironments around the protonated nucleotide (**Figure 3.4.4.1**). The current data set includes protonated adenine: the lead-dependent ribozyme in a A$^+$•C base pair, U6 internal stem loop (U6 ISL) in a A$^+$•C base pair, and the hairpin ribozyme in a protonated adenine interacting with a phosphate group, and protonated cytosine: the Hoogsteen DNA duplex of a G•C$^+$ base pair, and a protonated C$^+$GC base triplet in the beet western yellow virus (BWYV).

Table 3.4.4.1: Calculated pK$_a$ of various protonated nucleotides in reported RNA systems

System	Interaction	Exp pK$_a$	Calc pK$_a$
Lead-dependent Ribozyme	A$^+$•C	6.5	6.9
U6 Internal Stem-loop	A$^+$•C	6.5	8.1
Hairpin Ribozyme	A$^+$Phos	6.5	9.5
Hoogsteen DNA duplex	G•C$^+$	7.2	6.4
Beet western yellow virus	C$^+$GC	8.2	>13

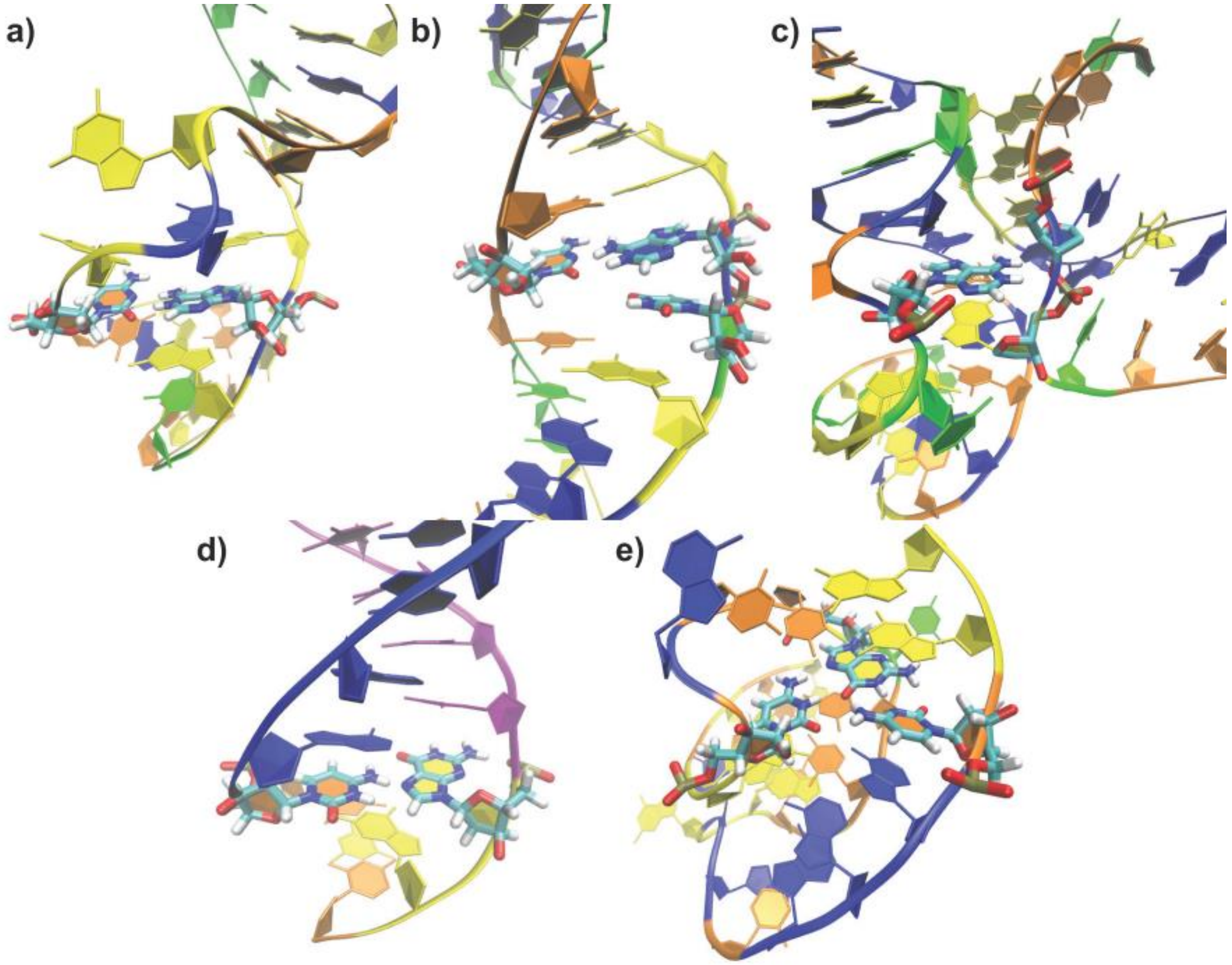


Figure 3.4.4.1: Ground state structure as obtained by X-ray crystallography or NMR studies of the 5 nucleic acid systems tested

A summary of the computed pK_a values from explicit solvent $CPHMD^{MS\lambda D}$ simulations is reported in **Table 3.4.4.1**. For adenine, we observed that our calculations yielded values of 6.9, 8.1, 9.5 for each RNA structure, relative to their experimental value of 6.5. Interestingly, we observed that there is an apparent deviation from experimental pK_a for the U6 ISL and the hairpin ribozyme even though the A^+•C base pair for the lead-dependent ribozyme is highly accurate. From available literature on the dynamics of the U6 ISL bulge, it is known that it experiences a pH-dependent conformational change where the U80 base adjacent to the $A79^+$•C67 base pair "flips out" when the pH is lowered to 5.7 (**Figure 3.4.4.2b**), while it remains stacked with A79 when the pH is raised to 7.4 (**Figure 3.4.4.2a**). Our prior work on nucleic acid $CPHMD^{MS\lambda D}$ simulations have also demonstrated that conformational dynamics and consequently the local electrostatic environment around titrating residues affect the accuracy of computed pK_a values, such as those in the GAAA tetraloop region.[113] When we performed the our calculations on the other flipped out U6 ISL structure, we observed that the pK_a is lowered to 4.8, and the lower and upper bound pK_a values of both structures encompass the experimental pK_a of 6.5 It should be noted that the experimental pK_a measured is a macroscopic or apparent pK_a value, which is obtained from a superposition of both structures, and therefore should not be expected to be reproduced by our pK_a calculations on a single conformation. Based on this precedence of how different conformations can contribute to and affect the macroscopic pK_a measured, for the hairpin ribozyme with the elevated pK_a of 9.5, it also suggests conformational dynamics may be at play as well. Alternatively, one may expect that the stronger interaction with a negatively charged phosphate group (as opposed to a neutral cytosine in the A^+•C base pair) should stabilize the protonated state of adenine in the hairpin ribozyme, thus stabilizing the protonated form even further and elevating its pK_a value beyond 6.5. For cytosine bases, we observed that the pK_a of the protonated Hoogsteen G•C$^+$ base pair is 6.4, which is close to the indirect/inferred pK_a measurements obtained from NMR relaxation dispersion spectroscopy[117] that was performed on a methylated variant that traps the base pair in the Hoogsteen conformation. For the BWYV, we note that the pK_a of Cyt is elevated above 13, which is much higher than the 8.2 value reported from experimental studies. An examination of the electrostatic environment between a G•C$^+$ base pair and a C$^+$GC base triplet suggests that the protonated cytosine should be stabilized to a greater extent in the base triple environment of BWYV, and therefore an upward pK_a shift is not unexpected. In addition, it has been postulated from experimental observations that a non-ground state structure may exist,[185] which would be consistent with a number of studies that demonstrate the conformational flexibility of RNA structures involved in pH-mediated activity.[21,35] Regardless of the deviation from experimental pK_a values, we suggest that our pK_a calculations are internally consistent based on the local electrostatic environment and its effect on stabilizing charged nucleobases, but paradoxically they do not correspond to the experimental measurements of apparent pK_a values. This apparent inconsistency will be reconciled in **Chapter 4.2**.

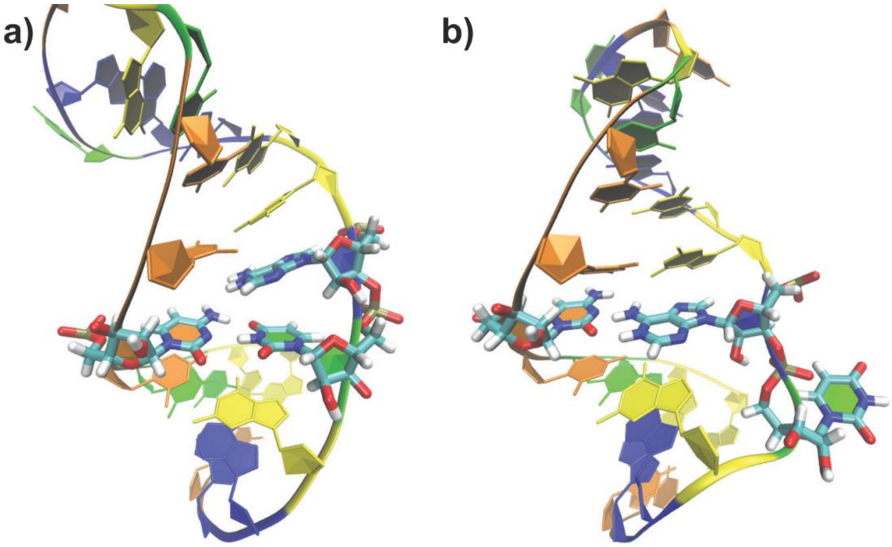


Figure 3.4.4.2: pH-dependent conformational change in the bulge region of U6 ISL adjusts the favorability of the A+C base pair interaction.

3.4.5. Conclusion: Expanded Explicit Solvent CPHMD$^{MS\lambda D}$ Simulations to include all Titratable Groups using an AMBER-compatible Force Field

In this chapter, we described the development of new parameters for nucleic acids in alternative protonation states for the AMBER force field. We tested a number of solvation models, basis sets and levels of theory for the QM calculations for deriving the RESP charges used to describe the partial charge distribution. We discovered that there is an apparent insensitivity of interaction energies and the resulting pK_a calculations of protonated A•C$^+$ base pairs to the specifics of the QM calculations. With the development of AMBER-compatible parameters, we then extended the explicit CPHMD$^{MS\lambda D}$ framework to titrate all the 5 major nucleobases present in both DNA and RNA, as well as additional functional groups, such as the backbone phosphate and 2'OH of the ribose sugar that have been implicated in pH-mediated RNA activity. Our pK_a calculations of protonated nucleotides across 5 different RNA structures indicate that the relative pK_a shifts are internally consistent based on the strength of the interactions that the protonated base has with its local microenvironment, but paradoxically is not always consistent with experimental pK_a measurements. Using the U6 ISL as a precedent, we demonstrate how different pH-triggered conformational changes can alter the microscopic pK_a of each conformation, and that the apparent pK_a measured is likely to be a superposition of the calculated microscopic pK_a values of these conformations. Based on the observation that the pK_a values of the protonated residues in the hairpin ribozyme and BWYV are elevated, we hypothesize that there may be additional conformational states, possibly transiently populated, that are involved in their activity at physiological pH.

Chapter 4: Using CPHMD$^{MS\lambda D}$ to Probe pH-mediated Transient States in Nucleic Acid Activity

4.1 Elucidating Transiently Populated Protonated Hoogsteen C•G^{+} Base Pairs in DNA Duplexes

Note: Chapter 4.1 was adapted from the following references.[117] *The entire chapter 4.1. contains significant contributions from Evgenia Nikolova, who was responsible for performing and analyzing all NMR experimental data. The results and discussion have been included in this dissertation for continuity and completeness.*

4.1.1 Hoogsteen Base Pairs in DNA and Implications in Biological Systems

Recent NMR studies using relaxation dispersion techniques[186,187] of A•T and G•C base pairs in duplex DNA indicate that they can transiently form Hoogsteen base pairs with populations in the range of 0.1–0.5% and lifetimes of 0.3–1.1 ms at pH ∼6.[49,52] Transition from Watson–Crick (WC) to Hoogsteen (HG) base pairs requires a 180° rotation of the purine base about the glycosidic bond and, therefore, a change in the base orientation from anti to syn conformation.[188] While A•T HG base pairs retain two hydrogen bonds (H-bonds) upon this conformational change, G•C HG base pairs retain only a single H-bond unless cytosine N3 becomes protonated to form a second stabilizing H-bond (**Figure 4.1.1.1a**).

To date, N3-protonated cytosine in a G•C^{+} HG base pair has only been directly observed by NMR for triplex DNA, where the protonation constant (or pKa) of cytosine N3 was shown to be elevated by more than 5 units for G•C^{+} HG[189] as compared to the value of ∼4.2 in free nucleotides.[142] However, the protonation state of cytosine N3 in G•C HG base pairs within duplex DNA has not been determined. The pK_a of free cytosine is far from neutrality (~4.2),[142] and the cytosine imino H3 proton cannot be directly visualized in crystal structures or by NMR measurements owing to rapid exchange with solvent. Indeed, the initial proposal that replication by human DNA polymerase ι (hPolι) proceeds via HG rather than WC pairing[190] was challenged on the grounds that at physiological pH, G•C would not exist as a stable HG base pair due to lack of protonation.[191] Although X-ray structures of duplex DNA bound to proteins, including hPolι (at pH ~6.5)[192] and TATA-binding protein (at pH ~6),[193] suggest that cytosine N3 and guanine N7 atoms are within H-bonding distance, protonation of cytosine N3 could not be unambiguously established. Determining the protonation state of cytosine N3 and its pK_a value becomes significantly more challenging in naked duplex DNA, where the HG base pairs exist only transiently in solution. In this study, we contributed computational methods to support NMR experiments to directly examine the pK_a of cytosine N3 in naked duplex DNA and relative stability of HG base pairs under physiological pH.

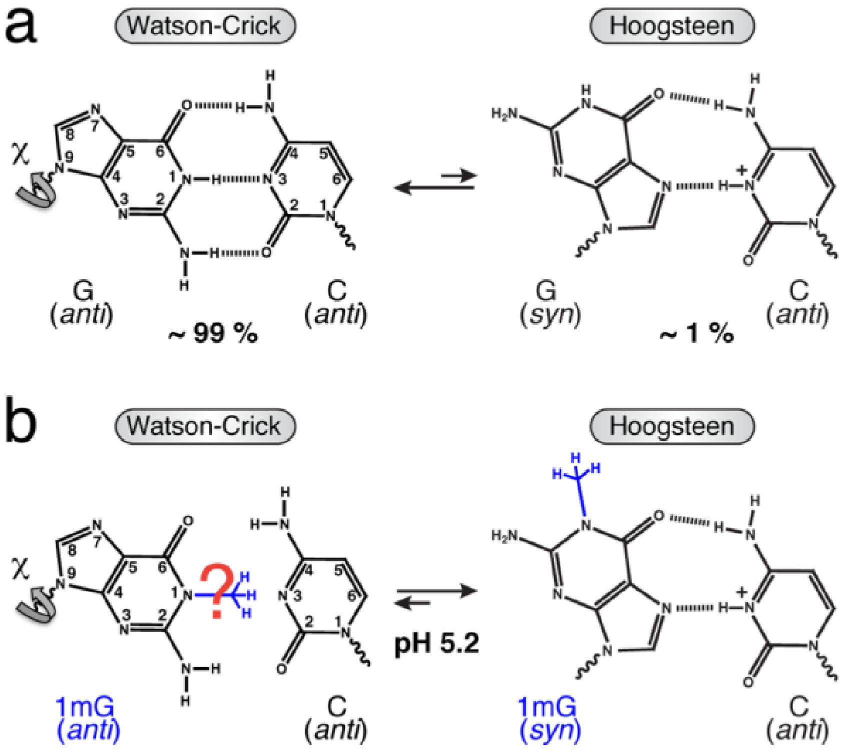


Figure 4.1.1.1: Schematic of the equilibrium between G•C WC and HG base pairs. (a) Transition from a ground-state WC to a transient-state HG base pair, with relative populations measured by NMR relaxation dispersion, requires an anti-to-syn rotation around the glycosidic bond (χ) and creates a stabilizing H-bond upon C N3 protonation. (b) Methylation at G N1 favors formation of a ground-state HG base pair at pH 5.2.

4.1.2 NMR Relaxation Dispersion Measurements of G•C HG Base Pairs

We previously showed that G·C HG base pairs can be trapped inside naked duplex DNA by installing a methyl group at the G imino nitrogen N1 position.[49] This N1-methylguanine (1mG) modification introduces a bulky substituent at the WC interface and precludes formation of the WC (G)N1H1···N3(C) H-bond, tipping the equilibrium toward the HG base pair at low pH (**Figure 4.1.2.1b**).[49] Based on chemical shift analysis, we showed that trapped HG base pairs have similar characteristics to their transient unmodified counterparts. We confirmed formation of the 1mG15·C10 HG base pair in A6-DNA 1mG10 at pH 5.2 based on observation of nuclear Overhauser effect (NOE) connectivity and proton/carbon chemical shift signatures that indicate a syn conformation for the 1mG10 base (**Figure 4.1.2.1a**).[49]

While the protonation state of cytosine could not be deduced directly in either transient or trapped HG base pairs, several indirect lines of evidence suggest that in both cases, the cytosine N3 is protonated to form a G•C$^+$ HG base pair. The 1mG10 modification resulted in significant chemical shift perturbations at the C15 base, which are consistent with N3 protonation. This includes an upfield shift of amino protons (∼2 ppm), which is a known characteristic of protonated G•C$^+$ HG base pairs in triplex DNA,[189] and a large downfield shift (∼2.3 ppm) in C15 C6, which is also expected upon N3 protonation based on density functional theory calculations.[49] Further evidence that these perturbations reflect cytosine N3 protonation comes from observation of only small chemical shift perturbations (<0.5 ppm) in the thymine residue when trapping an A•T HG base pair through N1-methylation of the adenine.[49] Finally, the population of the transient HG base pairs measured by NMR relaxation dispersion decreases more strongly with increasing pH for G•C versus A•T base pairs, and falls outside the limits of detection by relaxation dispersion at higher than neutral pH, as might be expected based on destabilization of the G·C HG base pair due to cytosine N3 deprotonation.[49]

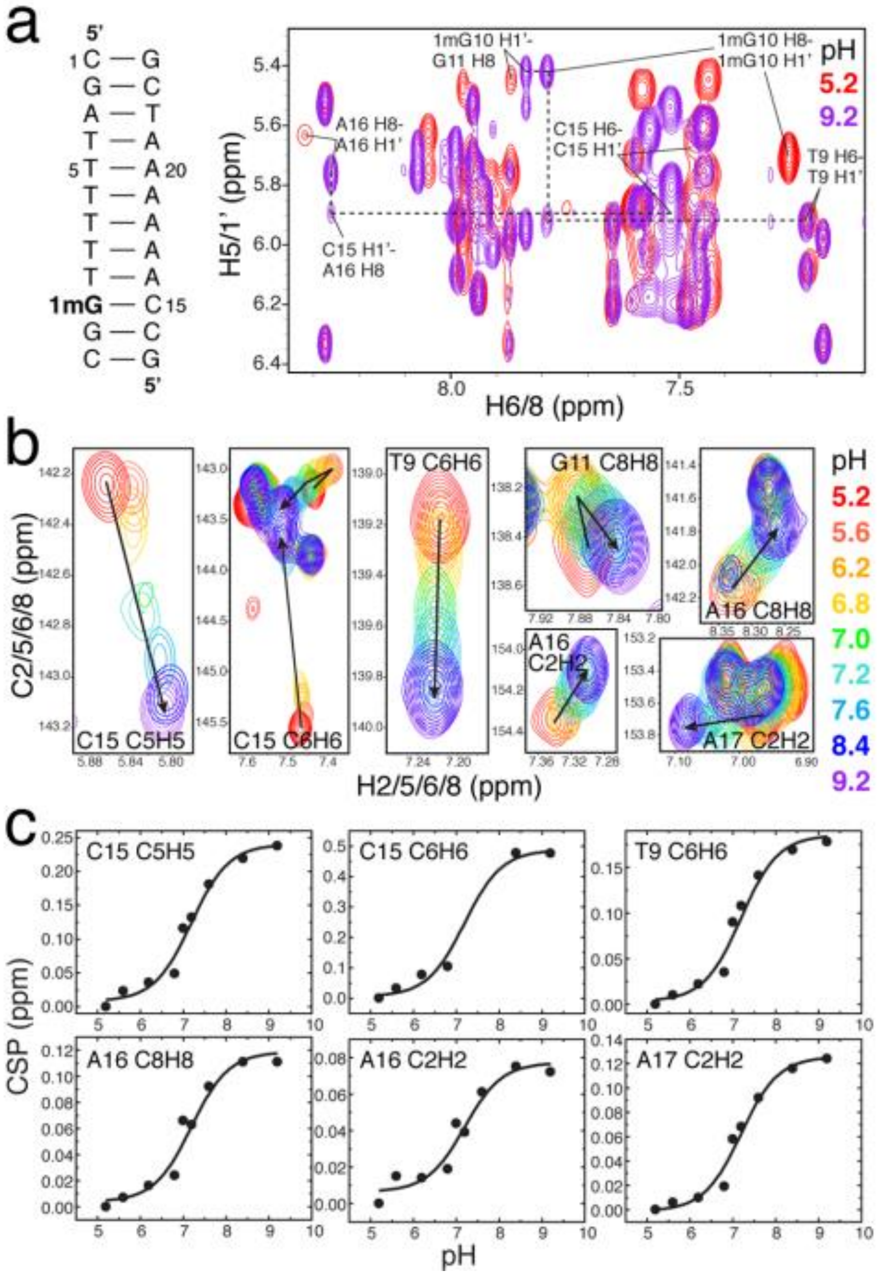


Figure 4.1.2.1: Estimating the pK a for cytosine N3 inside a trapped 1mG · C HG base pair. (a) 2D1H,1H NOESY spectra at pH 5.2 (red) and 9.2 (purple), suggesting a syn conformation at low pH versus an anti conformation at high pH for 1mG10 as well as enhanced conformational exchange and/or distortion for C15 and neighboring sites. (b) pH dependence of 2D1H,13C

HSQC spectra of unlabeled A 6 -DNA1mG10 showing large conformational changes at the 1mG10 · C15 and its two neighboring base pairs. (c) Corresponding chemical shift perturbations as a function of pH, showing global fitting of the observed pK_a ≈ 7.2 for the transition from a protonated G · C+ HG to a distorted WC * base pair

To further characterize the protonation state of C15 N3 in a G·C HG base pair, we measured natural abundance NMR 1H,13C-HSQC spectra for base and sugar resonances for the unlabeled A6-DNA1mG10 sample as a function of pH and monitored the chemical shift perturbations (CSP) at the modified base pair and adjacent sites (**Figure 4.1.2.1b**). We worked within a pH range (5.2–9.2) that minimally affects the structural stability of B-DNA and that causes little NMR spectral change in an unmodified A6-DNA. If the chemical shift perturbations observed at C15 upon guanine methylation under acidic conditions arise due to protonation of cytosine N3, increasing the pH should undo these effects and result in C15 chemical shifts that are similar to those observed in WC base pairs. Increasing the pH from 5.2 to 9.2 resulted in expected upfield CSPs for cytosine C6 and C5 that are consistent with deprotonation at the N3 position (**Figure 4.1.2.1b**). However, we also observed CSPs that are not expected based on N3 deprotonation and that suggest a pH-dependent conformational change. In particular, both the sugar C1′ and base C8 resonances of 1mG experience an upfield shift with increasing pH, resulting in carbon chemical shifts that are strongly indicative of an anti rather than syn nucleobases orientation, as expected for a WC-like geometry. This was supported by large changes in the NOESY cross-peaks at pH 9.2, including a much weaker 1mG10 H8–H1′ cross-peak and a stronger 1mG10 H8–H2′/2″ cross-peak than seen for the syn base at pH 5.2, but consistent with an anti base orientation (**Figure 4.1.2.1a**). We also observed a weak cross-peak between 1mG10 H8 and the 3′ neighboring T9 H1′, confirming that an anti/anti configuration in the sequentially stacked bases, with some structural distortion and/or enhanced dynamics at the 1mG residue (**Figure 4.1.2.1a**). Increasing the pH resulted in an unusual downfield CSP for C15

C1′ that suggests a change in sugar pucker toward the C3′-endo conformation. A structural and/or dynamic perturbation at C15 could also be inferred from a weaker cross-peak between C15 H1′ and the 3′ adjacent A16 H8 at pH 9.2 than normally observed in B-DNA (**Figure 4.1.2.1a**). These data suggest that, upon deprotonation of cytosine N3 at high pH, an HG base pair stabilized by a single H-bond is no longer energetically favorable as compared to a distorted WC-like geometry (WC*), which could be stabilized by at least one H-bond. Evidently, the 1mG modification does not fully trap the transient HG base pair at pH 5.2 but, rather, inverts the relative populations of the WC and HG species so that the WC* conformation now becomes the transient state. This is further supported by detectable line broadening at the 1mG10·C15 base pair observed at low pH. Such inversion of ground and excited states has previously been observed with targeted mutagenesis in proteins.[194]

4.1.3 CPHMD$^{MS\lambda D}$ Simulations of G•C HG Base Pairs

To obtain additional insights into the protonation equilibria, we performed constant pH molecular dynamics (CPHMD$^{MS\lambda D}$) simulations[111,112] on the HG G•C+ base pair and its 1mG analogue using the same NMR experimental conditions. As shown in **Figure 4.1.3.1a**, we calculated pK_{HG+} = 7.1 ± 0.1, where the major neutral HG conformation was stabilized by two weaker H-bonds (**Figure 4.1.3.1b**). Moreover, this pK_a prediction was not significantly altered by guanine N1-methylation (**Figure 4.1.3.1a**). Analysis of the H-bond lengths at pH 7 confirmed that an HG-like conformation was maintained throughout the simulations (data not shown). These results represent an independent estimate of pK_{HG+}, which is in line with the experimentally bounded pK_{HG+} value of at least 7.2 ± 0.1, and point to a nearly equal stability of the neutral and protonated species at physiological pH. As in the NMR experiments, the MD simulations may underestimate pK_{HG+} because polarization effects from the charged $G{\cdot}C^+$ base pair, which could strengthen these interactions, were not accounted for in the simulation parameters. In contrast, control simulations for a canonical WC base pair , where the protonated species featured a cytosine base shifted toward the major groove to accommodate a wobble configuration with two H-bonds (**Figure 4.1.4.1b**), yielded a much lower $\mathrm{p}K_a = 2.4 \pm 0.1$ that fits the large decrease expected for a helical WC base pair. Due to the lack of accurate structures for the protonated and neutral WC* states, identical simulations could not be carried out for the 1mG-modified WC* base pair.

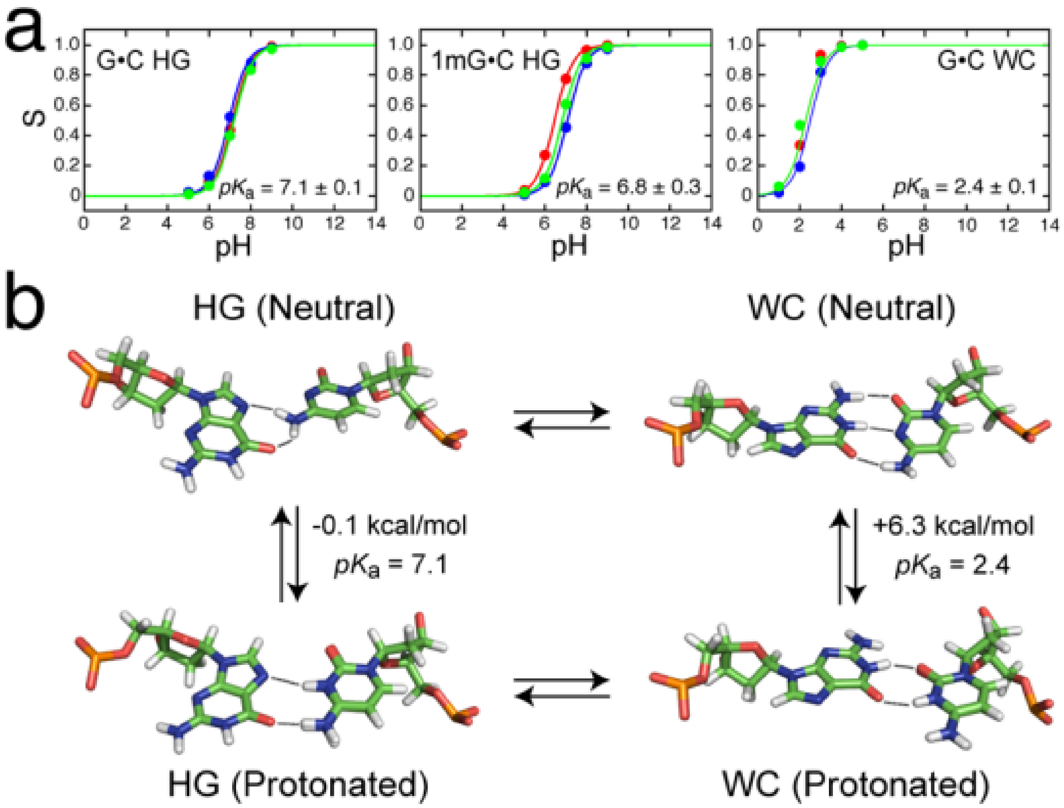


Figure 4.1.3.1: Constant pH MD simulations of WC and HG base pair protonation. (a) Titration curves obtained from three independent runs of single-site $\mathrm{CPHMD}^{\mathrm{MS}\lambda\mathrm{D}}$ simulations of a G · C HG base pair, its 1mG analogue, and a G · C WC base pair. (b) Corresponding structures for the neutral and protonated WC and HG base pairs and predicted free energy differences at pH 7, depicted in the context of the proposed four-state equilibrium.

To relate the above observations to transient HG base pairs, we measured relaxation dispersion data over the detectable pH range (4.3–6.8) to examine variations in the HG population (pB). Assuming that the neutral G•C HG base pair is significantly destabilized relative to its protonated counterpart, we would predict that, at pH > pK_a of cytosine N3 (≥7.2), G•C HG base pairs would fall outside the limit of detection by NMR dispersion. This would not be the case for A•T HG base pairs, whose populations should remain independent of pH. Indeed, this is what is observed: transient G•C+ HG base pairs are undetectable at pH 7.6, while A•T retains a pB ≈ 0.4%. By extrapolating the pH dependence of pB, we estimate a pB ≈ 0.02 to 0.002% for transient G•C+ HG base pairs at physiological pH 7–8. This is at least ~20-fold less abundant than for transient A•T HG base pairs, and this difference in abundance increases with metal ion concentration (data not shown). A comprehensive survey of X-ray structures also reveals a greater abundance of A•T as compared to G•C HG base pairs in duplex DNA (data not shown). Interestingly, we also observed an increase in pB with decreasing pH below 6, which is much more pronounced for G•C+ as compared to A•T base pairs. Fitting of pB as a function of pH yielded pKa,obs = 3.2 and 2.7 for G•C and A•T base pairs, respectively (see Supporting Information). This increase in pB with acidic pH arises primarily from an increase in the forward rate constant and could reflect acid-induced destabilization[195] of WC relative to HG states, possibly due to protonation of other groups. For G•C base pairs, this increase in pB could still be explained by cytosine N3 protonation in the context of a four-state equilibrium.

4.1.4 Conclusion: CPHMD$^{MS\lambda D}$ Simulations were used to Characterize Transiently Populated Hoogsteen G•C$^+$ Base Pairs in DNA

In this chapter, we have used both NMR and CPHMD$^{MS\lambda D}$ simulation studies, which both indicate that the pK_a of cytosine N3 is ~7.2, which is comparable to the pK_a of adenine N1 in A·C$^+$ mismatches.[39,196] Thus, transient G·C HG base pairs can significantly populate protonated over neutral species near biological pH, with potential implications in DNA recognition and binding by cellular factors. Moreover, we show that, at physiological pH, G·C base pairs containing N1-methyl-G damage exist as a nearly equal mixture of protonated HG^+ and distorted WC-like conformers that could be specifically recognized by DNA repair enzymes in search for damaged DNA.

4.2 The Role of Transient States in Hairpin Ribozyme Catalysis

Note: Chapter 4.2 was adapted from the following references.[116] The entire chapter 4.2 contains considerable contributions from Kamali Sripathi, who was involved in performing part of the simulations and data analysis.

4.2.1 The Hairpin Ribozyme: Embroiled in Controversy

Since the discovery of RNase P and the Tetrahymena group I intron, catalytic RNAs have been found to catalyze a variety of reactions and exist in a wide array of structural size and diversity.[197,198] The catalytic power of large ribonucleoprotein complexes such as the ribosome and spliceosome have been proven to be due to their RNA components.[197,198] The ribosome, one of the largest ribozymes, catalyzes a very specific reaction, the formation of peptide bonds, and it is distinct from the phosphoryl transfers of other ribozymes. The reactions catalyzed by both RNase P and the group I intron, the latter of which is considered a ribozyme of intermediate size,[199] have been shown to be metal-dependent. Smallest of all, the autolytic ribozymes, which include the Hepatitis delta virus (HDV) ribozyme,[11,200,201] the hairpin ribozyme,[202,203] the hammerhead ribozyme,[204] and the Varkud Satellite ribozyme,[205] catalyze site-specific cleavage of their own phosphodiester backbones primarily through the use of nucleobases in their alternative protonation states.[197-199]

In the context of the small autolytic ribozyme, the hairpin ribozyme is perhaps the best studied and characterized system. It facilitates site-specific autolytic cleavage into 2’3’ cyclic phosphate and 5’ hydroxyl termini, and adopts a variety of conformations with several discrete strategies to carry out site-specific cleavage. Extensive biochemical and structural data have indicated that the hairpin ribozyme employs general acid-base catalysis, which is believed to be facilitated by its own endogenous nucleobases. An alternative catalytic mechanism, via electronic stabilization, has been proposed by Fedor and co-workers.[12,40] However, more recent publications have demonstrated how existing data supporting the latter hypothesis can be re-interpreted in the context of general acid-base catalysis, further solidifying this mechanism as the consensus within the community.[206,207]

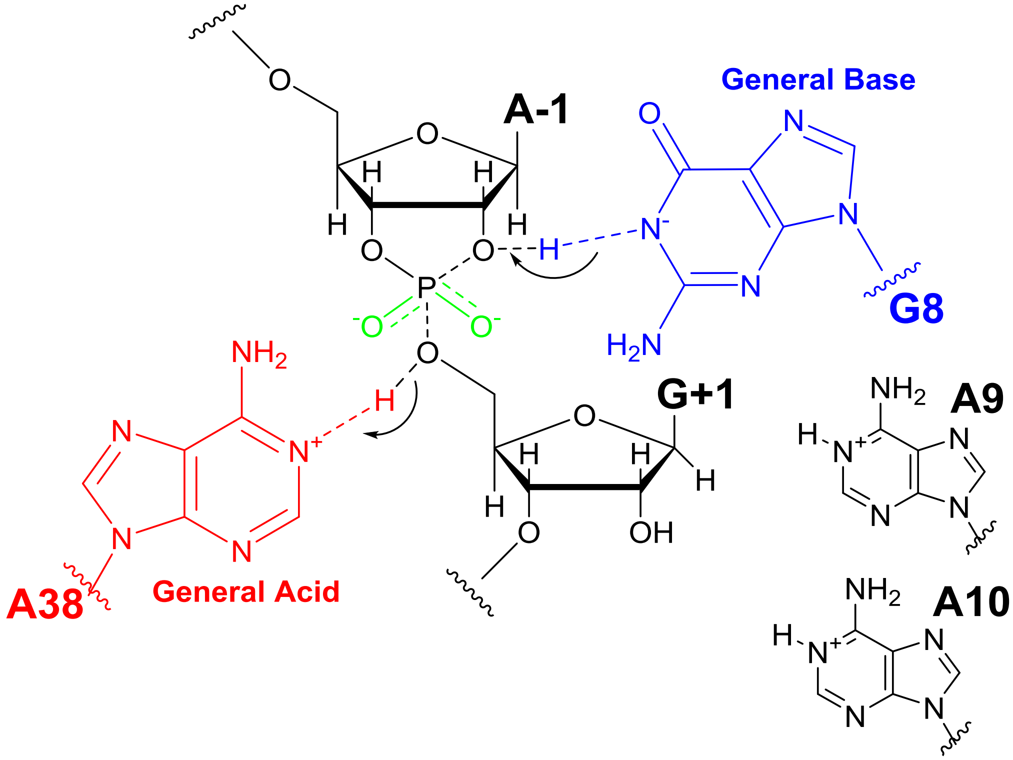


Figure 4.2.1.1: Proposed mechanism of the hairpin ribozyme A38-G8 general acid-base catalytic mechanism. Alternative candidates for the general acid and/or base, and supporting residues with residual effect have also been highlighted.[12,41,44,208209,210]

As to the specific details of the catalytic residues, there are several candidates for general acids and bases in the hairpin ribozyme active site. A38 in its protonated form has been implicated as the general acid,[12,41,44,208] as evidenced by its site-specific pK_a value of 6.5, which suggests that it can be partially protonated near physiological pH. As for the general base, G8 in its deprotonated form is a possible candidate, although its high pK_a of ~10.5 suggests that the population of $G8^-$ is going to be extremely low near physiological pH. The A38/G8 general acid/base catalytic model is further supported by mutational studies where the loss of A38 and

G8 leads to a 10^4 and 10^2 drop in catalytic activity respectively. In a seminal work by Bevilacqua, it was demonstrated how a general acid at pK_a ~6.5 and general base at pK_a ~10 can be used to reconstruct experimentally recorded pH rate profiles, and the A38/G8 general acid/base catalytic model is thus far the most consistent with the majority of experimental data.

While the A38/G8 general acid/base catalytic model is the most widely accepted model for the hairpin ribozyme, it is not fully consistent with all experimental observations. Notably, in a recent study involving thiol substitution and A38 abasic mutation by Lilley and co-workers, a "residual" catalytic effect at low pH was observed that cannot be explained by the A38/G8 model.[209] Over the years, a number of alternative models for the identity of the general acid and general base has been explored by the community. Early mutational studies have implicated residues A9 and A10, although the 10-fold drop in activity (relative to the 10^2 to 10^4 fold drop observed for A38 and G8), and their increased distance away from the active site likely relegates their role to electronic stabilization rather than actual catalytic participation. Early studies have also demonstrated that the hairpin ribozyme utilized an ion-independent mechanism of catalysis.[210,211] More recent studies using QM/MM methods from the groups of Gao, York and Otyepka have explored the possibility of non-bridging oxygens to act as proton shuttles,[212,213] although the pK_a of these phosphate groups in the hairpin ribozyme environment have yet to be measured or calculated in the literature.

In this chapter, we use pH-REX CPHMD$^{MS\lambda D}$ to examine the protonation equilibria of all residues implicated in the catalytic mechanism of the hairpin ribozyme from first principles, which includes G8, A9, A10, A38, non-bridging oxygens on the scissile phosphate, and the 2'OH group of the ribose sugar as illustrated in **Figure 4.2.1.1**. Based on our preliminary pK_a calculations using the hairpin ribozyme crystallographic structure, and the conformational flexibility we and others have observed in nucleic acids,[21,35] we will map out various conformational states, including transiently populated states, in order to examine how the local environment around various titrating sites affect their protonation equilibrium, and how pH-dependent transient states affect the overall catalytic mechanism of the hairpin ribozyme.

4.2.2 Initial pK_a calculations using CPHMD$^{MS\lambda D}$ simulations

Our earlier pK_a calculations of several protonated adenine species in different RNA systems yielded the interesting observation that the microscopic pK_a of A38 in the hairpin ribozyme is elevated to ~9.5, which is not consistent with the site-specific pK_a of 6.5 as measured from Raman crystallography and pH-sensitive fluorescent nucleobases analogs experiments.[41,44] We note that there is precedence of pK_a variation of adenine residues even within similar A^+C base pair environment, as reported by Bevilacqua and co-workers, where they showed how flanking bases can affect the pK_a of A^+C base pairs.[214] In addition, extensive studies on the U6 ISL has confirmed that it undergoes a pH-dependent conformational change, which we have shown to exhibit distinct microscopic pKa. Furthermore, in dynamic systems such as the U6 ISL, caution must be exercised when interpreting structural data, as the initial structural studies performed at physiologically relevant pH conditions were incorrect because NMR signals from the two conformations at high and low pH were not deconvoluted properly.[35,215,216]

To test the hypothesis that conformational changes may be influencing the measured pK_a of A38, we mapped out possible conformations along the reaction pathway of A38, using the A38(N1)…G+1(P) distance as a proxy for the reaction coordinate, which represents the approach of a protonated A38 to the center of mass of the cleavage site (scissile phosphate). Using WExplore, a hierarchally balanced weighted ensemble sampling technique, we identified a major population centered at about ~ 5 Å, and this structure is similar to that observed in crystallographic studies (**Figure 4.2.2.1a**). In addition, another minor population at a distance of ~ 9 Å was also identified. In this minor population, which we will term the relaxed state, the increased distance between A38 and the negatively charged active site should decrease the electrostatic stabilization on a protonated A38, which will manifest as a lower pK_a. When we incorporated both major and minor states into pH-REX CPHMD$^{MS\lambda D}$ simulations, the pK_a of A38 decreased from ~9 to ~7 (**Figure 4.2.2.1b**), bringing it closer to the experimentally recorded pK_a of 6.5. These findings suggest that, on the timescale of experimental measurements, the hairpin ribozyme fluctuates between the ground state crystallographic structure and a transiently populated relaxed structure. Furthermore, it also implies that the measured pK_a of A38 does not correspond to the crystallographic structure, as the crystal structure appears to have a much higher microscopic pK_a than the apparent value, which would be obtained as a superposition of the microscopic pK_a values of both the ground and relaxed structure.

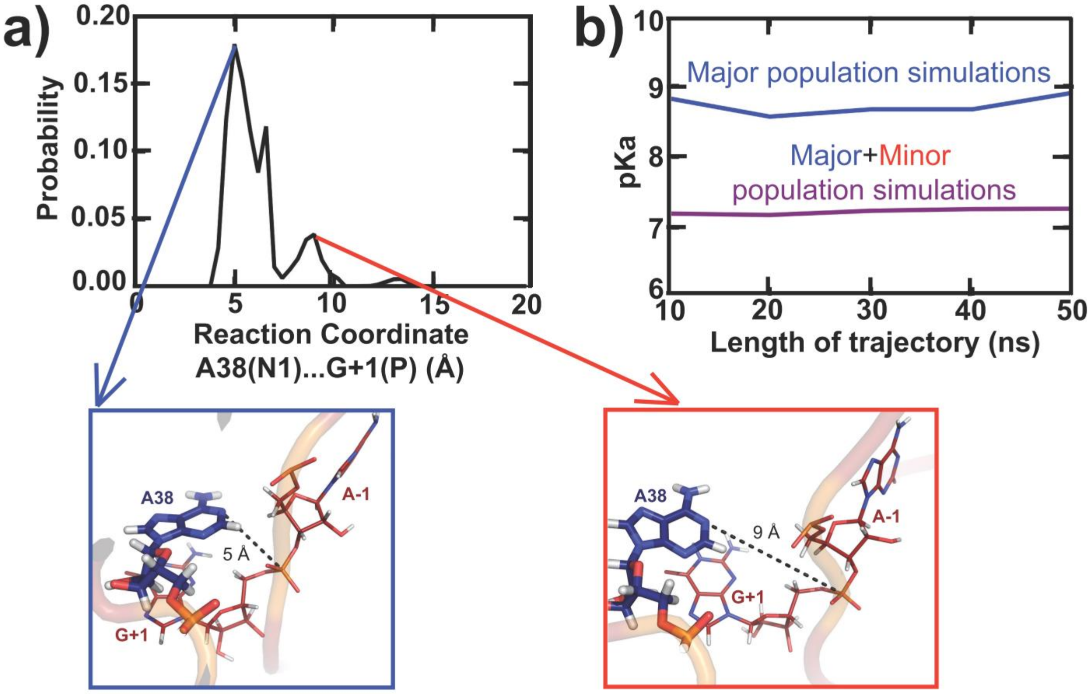


Figure 4.2.2.1: 1D WExplore sampling along the reaction coordinate identified a transiently populated relaxed conformation that when incorporated into CPHMD simulations better agreed with experimental data.

4.2.3 Identifying Transient Conformational States Involved in Catalysis

To fully elucidate the role of these transiently populated relaxed structures, and to clarify the nature of various residues and functional groups implicated in the catalytic mechanism of the hairpin ribozyme, we used an extensive set of WExplore simulations to exhaustively map out the conformations along two reaction coordinates: (i) the A38(N1)…G+1(P) distance, and (ii) the G8(N1)…G+1(P) distance. In addition, 8 permutations of protonation states of A38, G8 and the 2'OH functional group were subjected to 2D WExplore sampling (**Table 4.2.3.1**). In the default WExplore simulation setup, the 2'OH retains its canonical protonated state. Low, medium, and high pH conditions were simulated by fixing the protonation states of A38 and G8 at their expected values as illustrated in **Table 4.2.3.1**. In addition, a transient state that corresponds to the $A38H^+$ and $G8^-$ protonation states, which represent the active species in the reaction mechanism, was included even though the population of this state is likely going to be extremely low. Apart from the 4 default WExplore simulations, an additional 4 alternative WExplore simulations were performed, where the 2'OH was adjusted to its negatively charged deprotonated state, which better represents a later stage of the reaction pathway.

Table 4.2.3.1: Permutation of fixed protonation states used in 2D WExplore sampling

Permutation	2'OH	A38	G8
def_transient	2'OH	A38H(+)	G8(-)
def_high	2'OH	A38	G8(-)
def_med	2'OH	A38	G8H
def_low	2'OH	A38H(+)	G8H
alt_transient	2'O(-)	A38H(+)	G8(-)
alt_high	2'O(-)	A38	G8(-)
alt_med	2'O(-)	A38	G8H
alt_low	2'O(-)	A38H(+)	G8H

From the collective results of the 8 WEXplore simulations, the identified conformations were clustered using the RMSD of all heavy atoms near the active site, which includes residues

G8, A38, and the two residues flanking the scissile phosphate (G1 and A-1). A total of 11 clusters were identified, and their medoid structures were extracted. Clusters that had G8 or A38 greater than 10 Å away from the center of the active site were excluded from further analysis, as in these structures partial unfolding of the active site was observed. The resulting 5 clusters contributed to 95.5% of all conformations sampled in the 8 WExplore simulations, which ensures that the loss of data from excluding the partially unfolded structures should have little effect on the overall analysis. Separate pH-REX CPHMD$^{MS\lambda D}$ simulations were initiated from the medoid of each of the 5 clusters. The resulting pK_a values, which should correspond to the microscopic pK_a are summarized in **Table 4.2.3.2**.

Table 4.2.3.2: Microscopic pK_a calculated from pH-REX CPHMD$^{MS\lambda D}$ simulations of the 5 dominant clusters as identified from 2D WExplore sampling.

Cluster	G8	A9	A10	A38	Phos	2'OH	Weight
0	13.0	2.0	2.0	13.0	2.0	9.4	0.3642
4	6.0	2.7	7.1	9.5	5.0	11.3	0.0953
5	7.4	2.0	2.3	2.9	6.3	11.7	0.0938
7	9.3	3.1	5.5	3.6	5.9	11.2	0.2204
9	10.7	4.4	3.1	5.6	5.9	5.8	0.1817

We observed that the microscopic pK_a of A38 follows a bimodal distribution that encompasses the apparent pK_a of 6.5 measured, with 3 clusters with a pK_a of less than ~6.5 and 2 clusters with a pK_a higher than ~6.5. This observation is consistent with our earlier 1D WExplore sampling and reinforces our hypothesis that the hairpin ribozyme undergoes considerable conformational fluctuations between two distinct microenvironments around A38, which alternately favor and disfavor protonation. Similarly, the pK_a of G8 shows a bimodal distribution about the apparent measured pK_a of ~10.6, with 3 clusters possessing a lower pKa, and 2 clusters having a higher pKa. For the 2'OH group, the pK_a was downshifted from its reference value of 13.1 to about ~11 in most clusters, although it drops to as low as 5.8 in one simulation. This downshift in pK_a value is not unexpected since the proton of the 2'OH is removed during the first step of the reaction mechanism. As for residues A9 and A10, it was noted that the pK_a of A9 is near its reference pK_a of 3.5 in all clusters, however for A10, the pK_a is elevated to between ~5.5 and ~7 in 2 clusters. Lastly, the non-bridging oxygens on the scissile phosphate are observed to have a pK_a that is upshifted from 1.3 to ~5 to 6 in all but the first cluster.

To reconcile our simulation results on the microscopic pK_a of the 5 clusters with experimentally inferred apparent pK_a values, we derived a relationship between microscopic pKa, macroscopic apparent pK_a and the free energy difference (or weights) between the various conformations. From each permutation of 2D WExplore simulations, the weights of the 11 clusters can be obtained. In order to consolidate information from all 8 permutations of protonation states (see **Table 4.2.3.1**), we averaged the weights across all 8 simulations, and used **equation 2.2.5.1** to calculate the apparent pKa. The results as summarized in **Table 4.2.3.3** indicate that the calculated apparent pK_a of G8 is within 0.5 pK_a units from the experimentally suggested value of 10.6. Based on pK_a considerations alone, the calculated apparent pK_a of A10 at 6.1 and the backbone phosphate at 5.9 would suggest that these residues may play a role in the catalytic mechanism as well. Interestingly, we observed that the apparent pK_a of A38 is much higher than anticipated at 10.5. We attribute this discrepancy relative to our prior results by noting that cluster 0 had a pK_a >13 for A38, which is anomalously high. To determine if the additional 6 excluded clusters would have any effect on the predicted apparent pK_a, we recalculated the apparent pK_a using all states, and as shown in **Table 4.2.3.3**, the values are extremely similar to those obtained using 5 clusters.

Table 4.2.3.3: Reconstructed apparent pK_a using microscopic pK_a from CPHMD simulations using average weights and modified weights targeted to experimental pK_a of 6.5 and 10.6.

Residue	Exp pKa	Apparent Calc pK_a (11-states)	Apparent Calc pK_a (5-states)
G8	10.6	10.3	10.8
A9	-	3.9	4.2
A10	-	5.6	6.1
A38	6.5	10.1	10.5
Phos	-	5.8	5.9
2'OH	-	10.4	10.9

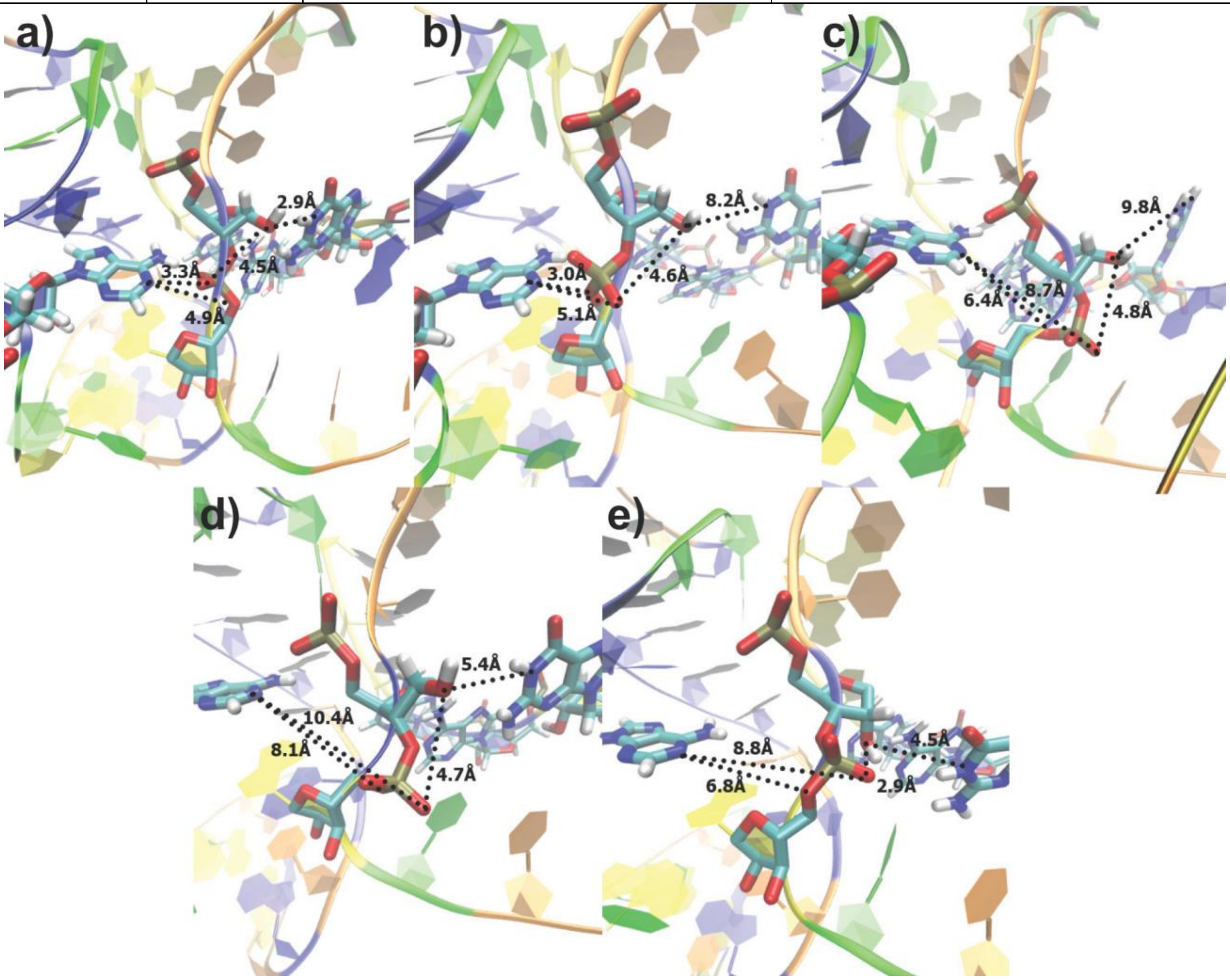


Figure 4.2.3.1: Representative conformation of each cluster as identified from 2D Explore sampling. Relevant distances to the approach of the reaction coordinate, notably A38(N1)…O5', A38(N1)…O2P, O2P…O2' and G8(N1)…O2' are illustrated in the figures.

Next, we examined the structural features of each of the 5 clusters identified, notably the distance between A38(N1)…O5', to describe the proton transfer between A38 and the active site, A38(N1)…O2P, to describe the proton transfer between A38 and the backbone phosphate, O2P…O2' to describe proton transfer between the backbone phosphate and the 2'OH group, and and G8(N1)…O2' to describe the proton transfer between G8 and the active site. Cluster 0 (**Figure 4.2.3.1a**) was populated 36.4% in all 8 2D WExplore simulations, and based on RMSD, it most resembles the ground state crystallographic structure. However, we note that in this medoid structure, A38 is positioned in-line (3.3 Å) for a proton transfer to the non-bridging oxygen on the scissile phosphate, which would be a catalytically incompetent structure. In addition, the high charge on the non-bridging oxygen would explain its elevated microscopic pK_a of >13. Further analysis of the structures within cluster 0 indicates that it may not be representative of the local environment around A38, and it can be positioned next to O5' or O2P, which suggests that this cluster would need to be further divided into sub-clusters to deconvolute mild structural differences that can have a substantially large effect on the microscopic pK_a calculated. As for residue G8, it is positioned in-line for a proton transfer (2.9 Å) to the O2' oxygen on the ribose sugar. Cluster 4 (**Figure 4.2.3.1b**) is populated 9.5% of the time, and may be viewed as a hybrid ground/transient conformation. Specifically, A38 is in the proper catalytically competent conformation, in-line (3 Å) for a proton transfer to the O5' oxygen on the scissile phosphate. This conformation around A38 corresponds to the crystallographic structure, and its pK_a is 9.5, which is consistent with our earlier calculations, and it also highlights the sensitivity of A38 with its interacting atom. In this cluster, the pK_a of A10 is also upshifted to 7.1, but it is not close enough to the active site to participate catalytically (>6 Å), and instead we suggest that it provides electronic stabilization of intermediate. G8 is positioned 8.2 Å away from the 2'OH group, and has a reduced pK_a of 6.0, which is more representative of a transiently populated relaxed structure. Cluster 5 (**Figure 4.2.3.1c**) is populated 9.4% of the time, and may be viewed as a transient relaxed structure, where both A38 and G8 are positioned away from the reaction site at 6.4 Å and 9.8 Å respectively. Not surprisingly, the microscopic pK_a of both residues are lowered to 7.4 and 2.9 respectively. It is interesting that in this fully relaxed conformation, where both residues are away from the active site, the phosphate group pK_a reaches 6.3, which is the highest observed amongst all the states we simulated. This suggests that the phosphate group may participate in the proton transfer, in situations where the general base and acid are absent from the active site, for example in an A38 abasic mutant. Cluster 7 (**Figure 4.2.3.1d**), which is populated 22% of the time, has A38 in a relaxed conformation (8.1 Å), but G8 is positioned at a pre-catalytic position at 5.4 Å away from the O2' atom. Here, G8 has a calculated pK_a of 9.3 and the pK_a of A10 is also elevated to 5.5. Lastly, cluster 9 (**Figure 4.2.3.1e**), which is populated 18% of the time, represents a possible catalytically competent state for the non-bridging oxygens on the phosphate group. Here, the O2P non-bridging oxygen is positioned in-line (2.9 Å) for a proton transfer to the O2' atom on the 2'OH group, and it has an elevated pK_a of 5.9. Notably, A38 is relaxed from the active site (6.8 Å) with a lowered pK_a of 5.6, and G8 is pre-catalytic at 4.5 Å away with a pK_a of 10.6.

4.2.4 Proposed Dual-Pathway Catalytic Mechanism

With the microscopic pK_a of all residues in the titrating site determined, and with the structural analysis of all major conformation states (including transiently populated states) in hand, we propose a dual-pathway catalytic mechanism to reconcile the disparate experimental observations made in the literature. Consistent with the consensus in the community that the hairpin ribozyme catalysis proceeds predominantly through a general acid/base catalytic mechanism involving A38 and G8 as illustrated in **Figure 4.2.4.1**. Our microscopic pK_a calculations are able to reconstruct the experimental apparent pK_a of A38 and G8 reasonably well, with the exception of A38, which we believe needs to be further deconvoluted particularly for cluster 0, although we note that the apparent pK_a is reproduced using pH-REX CPHMD$^{MS\lambda D}$

in our earlier 1D WExplore studies. A key finding from our simulations is the elevated pK_a of the non-bridging oxygens of the scissile phosphate, and the identification of a transient state where it is positioned for a proton transfer to the 2'OH group. Prior QM/MM studies have demonstrated the feasibility of the backbone phosphate role in catalysis, at least in terms of its energetics, given that the A38/G8 pathway compared to the Phos/Phos pathway have similar activation energies. However, the low population of a catalytically competent structure for the backbone phosphate and its lower pK_a indicate that the population of competent structures will be lower than that of the A38/G8 pathway. This indicates that the backbone phosphate can participate in the catalytic mechanism, but perhaps as a shadow pathway that is not the dominant pathway, but nevertheless may contribute under specific circumstances (**Figure 4.2.4.1**).

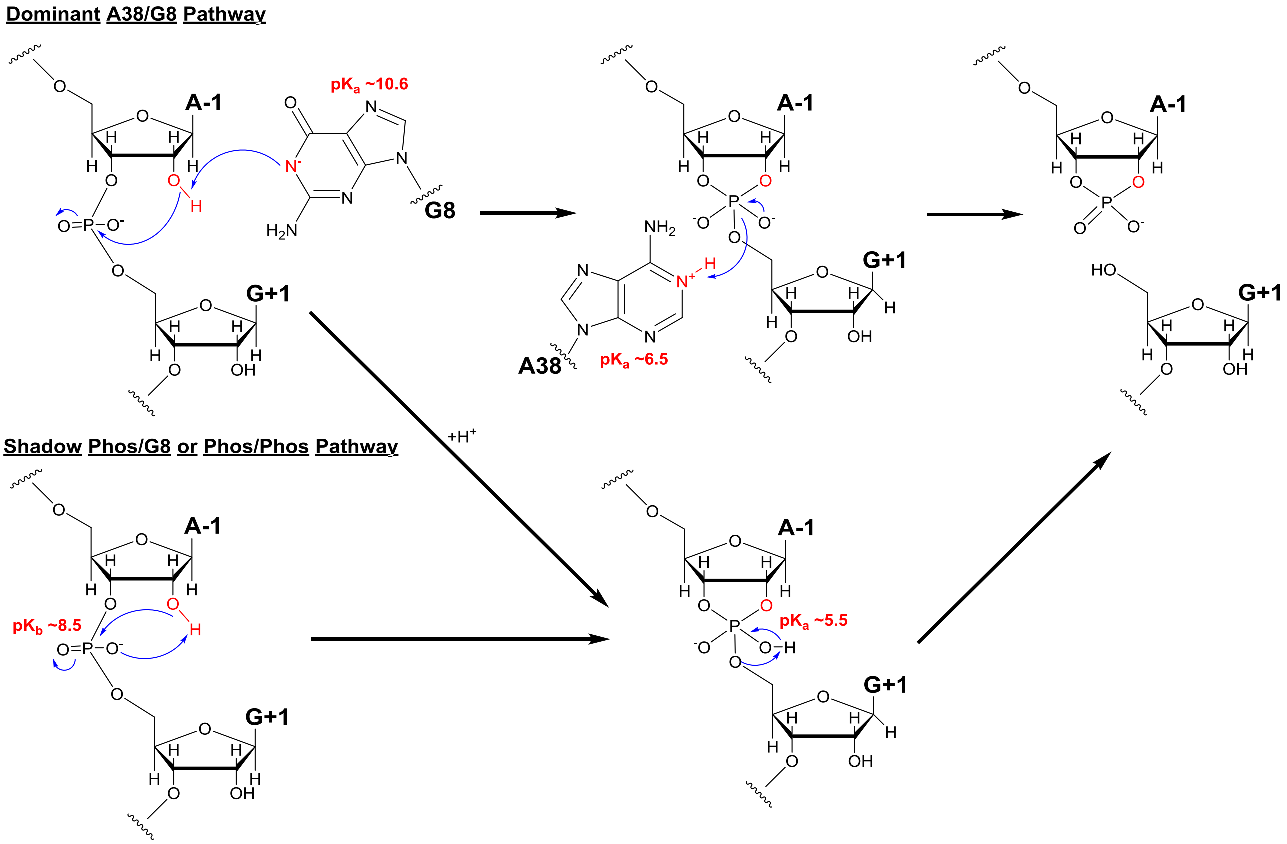


Figure 4.2.4.1: Proposed dual-pathway of the hairpin ribozyme catalysis.

Invoking the shadow dual-channel pathway explanation helps rationalize the differential effect on activity for mutational studies of A38 compared to G8, where loss of G8 led to a 2-fold smaller decrease in activity compared to A38. We can further rationalize this explanation by noting that the phosphate, when acting as a base, would have a pK_a of 14 – 5.9 = 8.1, which suggests that the hairpin ribozyme can use the non-bridging oxygens as the general base to compensate for the loss of G8 in the active site. Another situation where the shadow pathway may play a larger role in catalysis is when A38 is removed and/or at low pH conditions. Based on apparent pK_a values of A38 compared to the non-bridging oxygens on the scissile phosphate, the population of protonated species at physiological pH is expected to favor A38 by 10 to 100-fold. In the abasic A38 mutants constructed by Lilley and co-workers, a residual catalytic effect was observed under pH 6, which would be the condition at which the population of protonated phosphate will become sufficient to participate in the catalytic mechanism. As the role of the backbone phosphate is critical in this proposed dual-pathway mechanism, further experimental studies centered about this functional group, for example, substituting phosphate to thiol-phosphate, which has a different reference pK_a value, could result in observable differences in both the wild-type and the abasic A38 mutant. Lastly, our simulations have suggested a protonated A10 can exist in some transient states, although it remains too far from the active site to participate in catalysis. However, in the A38 abasic mutant, which would create a cavity near the active site, it is plausible that the A10 residue can position itself and serve the role as the general acid.

4.2.5 Conclusion: pH-mediated Transient States Identified from CPHMD$^{MS\lambda D}$ Simulations Propose a Dual-Pathway Mechanism of Hairpin Ribozyme Catalysis

In this chapter, we have used a combination of CPHMD$^{MS\lambda D}$ simulations augmented with WExplore enhanced sampling techniques to examine the catalytic mechanism of the hairpin ribozyme, by mapping out various conformational states visited by the hairpin ribozyme, and calculating the microscopic pK_a of each state. Notably, the discovery of pH-mediated transient states, particularly one that has an upshifted pK_a of the non-bridging oxygens of the scissile phosphate, led to the proposal of a dual pathway of the hairpin ribozyme catalysis: (i) a dominant catalytic pathway involving A38/G8 as the general acid-base, which is the consensus model in the field, and (ii) a shadow catalytic pathway involving the non-bridging oxygens on the backbone phosphate. This dual pathway mechanism was able to reconcile several puzzling observations, including the differential effects of mutational studies of A38 and G8 on the catalytic rate, and seemingly inconsistent experimental observations, including the residual catalytic effect observed in an abasic A38 mutant under low pH conditions. Furthermore, we have also identified that the ground state crystallographic structure, which best represents the catalytically active state of the hairpin ribozyme, has a pK_a of A38 that does not correspond to the experimental apparent pK_a. This is because the hairpin ribozyme fluctuates been the dominant crystallographic structure and several transiently populated relaxed state. In the context of the conformationally flexible RNA systems where pH-dependent transient states may exist, and owing to the challenge of deconvoluting and interpreting the pH effects from multiple transient states, our findings call for caution in using only ground state structures to interpret pH-mediated mechanisms.

Chapter 5: Using CPHMD$^{MS\lambda D}$ to Probe pH-mediated Transient States in Protein Activity

5.1 The Role of Transient States in Buried Ionizable Groups of Staphylococcal Nuclease Mutants

Note: Chapter 5.1 was adapted from the following references.[119]

5.1.1 Buried Ionizable Groups and their Broader Applications

The distinctive feature of explicit solvent CPHMD$^{MS\lambda D}$ compared to earlier generation implemented with implicit solvent models is the ability to model discrete ions and water, which would be applicable to systems that undergo partial desolvation such as in membrane proteins and ion channels. As a prelude to this goal, we investigated the performance of CPHMD$^{MS\lambda D}$ simulations on a general class of buried ionizable groups, in a series of staphylococcal nuclease mutants. Due to the charged nature of ionizable groups, the majority of them are expressed near the protein surface, and the bulk of our prior work on proteins (**Chapter 3.3**) have indicated that the CPHMD$^{MS\lambda D}$ framework is highly robust when dealing with such residues. However, the exception remains if dealing with conformationally "locked" residues such as those involved in salt bridges, or residues in systems such as membrane proteins where they can be located in more hydrophobic environments.

We focus our investigation on modeling the pH-dependent dynamics of a series of engineered mutants of staphylococcal nuclease (SNase) with Lys residues buried in the hydrophobic core[173] using the recently developed explicit solvent CPHMD$^{MS\lambda D}$ framework for proteins[114] While the effectiveness of CPHMD has been demonstrated on numerous systems,[94,95,97,98,217] almost all work reported to date was based on an implicit solvent model.[70,79,89] Applications of explicit solvent CPHMD on biomolecules thus far only have a few reported successes.[112-114,117,218] More importantly, none of the existing work has attempted a comprehensive investigation of buried ionizable residues, where, we hypothesize, the contrast in electrostatic environment between a hydrophobic pocket and a solvent-exposed environment will provide a key driving force for pH-mediated conformational fluctuations and, possibly, formation of transient states.

5.1.2 pK_a calculations using CPHMD$^{MS\lambda D}$ simulations

As shown in **Figure 5.1.2.1a**, the mutants for this study comprise a set of diverse mutation sites, with residues varying in the magnitude of pK_a shifts. In addition, experiments have demonstrated that for the mutants that we selected the titration of the internal Lys is decoupled from the ionization of Asp and Glu residues.[173] Therefore, to facilitate convergence within a reasonable time, we titrated only the buried Lys and performed an initial check of the accuracy of our simulations, by calculating its pK_a value, using the crystallographic structures as the starting models. As summarized in **Figure 5.1.2.1b**, most of the buried residues have a predicted $pK_a < 1$, which is shifted by nearly 10 pK_a units from the reference value of 10.4. pK_a calculations on the V66D and V66E mutants also produced a predicted pK_a of more than 14, which is dramatically shifted from their reference value of 4.0 and 4.4, respectively. In both cases, the pK_a shifted towards the direction stabilizing the neutral state of each titrating species (upward – for Asp/Glu, downward – for Lys). The dramatic pK_a shifts observed in our simulations mirror the experimental spectrophotometric measurements of Brønsted acids and bases, where a similar pK_a shift of 10-20 units was recorded when moving from an aqueous to an organic environment.[219,220] For example, acetic acid, which shares the same functional group with Asp and Glu, has its pK_a value shifted from 4.8 to ~23 in water (ε=80) vs. acetonitrile (ε=37).[220] Even though the dielectric constant of a typical hydrophobic pocket in SNase, which ranges from 4 to 20,[221,222] is lower than that of acetonitrile, such dramatic pK_a shifts of more than 10 units have never been observed in the protein. In fact, some of the largest pK_a shifts for Asp, Glu and Lys reported in the literature are perturbed by a mere 5 units from their reference values.[173,223] Thus, there is an inconsistency in the magnitude of experimental pK_a shifts observed in SNase compared to historical experimental pK_a shifts observed in organic solvents, even though both environments have a similar low dielectric value, which indicates that additional factors have to account for the apparent inconsistency.

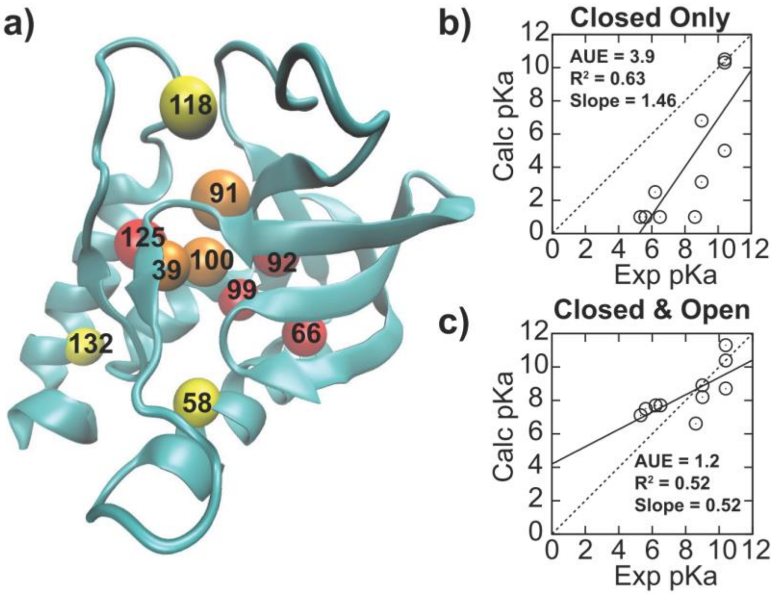


Figure 5.1.2.1: (a) Distribution of internal Lys residues of SNase mutants, color-coded depending on the pK_a shift: not shifted (yellow), shifted by 1-2 units (orange), and shifted by >2 units (red). Comparison between experimental and calculated pK_a values from explicit solvent

Strong coupling between pK_a and conformational sampling has been previously reported,[123,224-226] and, while the importance of conformational sampling have been observed in some SNase mutants, no connection to experimental pK_a values has been reported.[227] Based on these observations, we performed an additional series of extended-run explicit solvent pH-REX CPHMD$^{MS\lambda D}$ simulations on the structures pre-equilibrated for 50 ns at high and low pH.

Table 5.1.2.1: Calculated pK_a of SNase Lys mutants obtained from explicit solvent pH-compared to experimental pK_a values for all Lys mutants with highly shifted pK_a values.

Variant	Exp pK_a	CHARMM27 (Closed Only)	CHARMM27 (Closed & Open)	CHARMM36 (Closed & Open)
I92K	5.3	< 1.0	7.1	3.8
V66K	5.6	< 1.0	7.5	5.5
L125K	6.2	2.5	7.7	5.5
V99K	6.5	< 1.0	7.7	5
N100K	8.6	< 1.0	6.6	5.7
V39K	9.0	3.1	8.2	6.5
Y91K	9.0	6.8	8.9	6.7
A58K	10.4	5.0	8.7	9.9
N118K	10.4	10.5	10.4	10.4
A132K	10.4	10.3	11.3	11.2
AUE		**3.9**	**1.2**	**1.3**
R^2		**0.63**	**0.52**	**0.78**
Slope		**1.46**	**0.52**	**1.09**

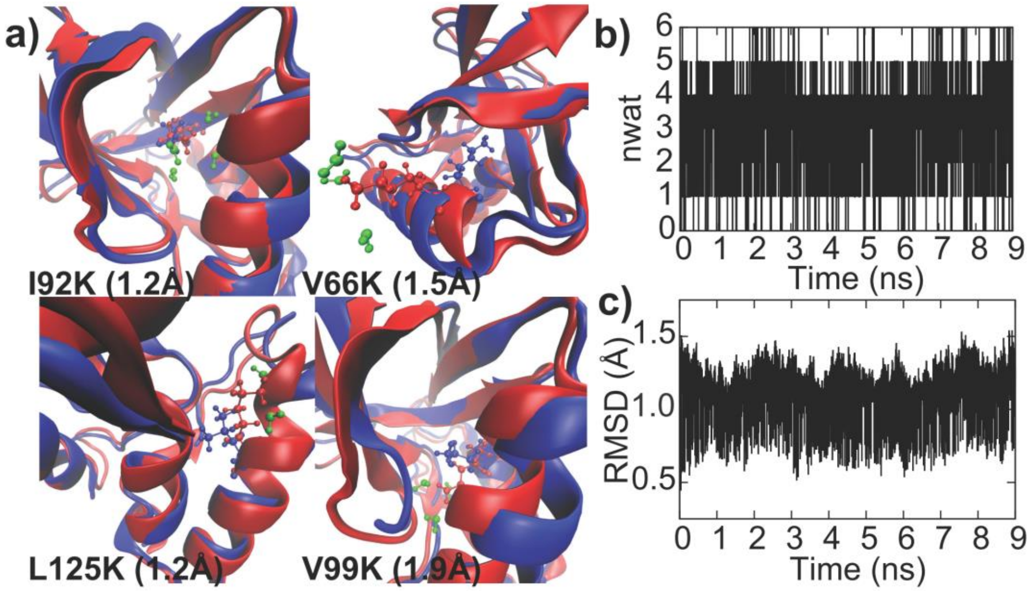


Figure 5.1.2.2: (a) Structures of the four most highly shifted Lys mutants at high pH “closed” conformation (in blue) and low pH “open” conformation (in red), with the backbone RMSD between both structures denoted in parentheses. Water molecules within 3 Å of the protonation site at low pH (in green) are included; no water molecules were observed at high pH. Using V66K, we show the conversion between the two states at external pH close to the calculated pK_a in the time evolution of (b) number of waters within 3 Å of the protonating site and (c) backbone RSMD relative to the X-ray structure.

The findings, summarized in **Figure 5.1.2.1c** and **Table 5.1.2.1,** demonstrate significantly improved results with the averaged unsigned error (AUE) reduced from 3.9 to 1.2 pK_a units. We observed that the longer equilibration allowed the sampling of "open" solvated structures that were critical for reproducing the experimental pK_a values reported for all Lys mutants with highly shifted pK_a values. This is in contrast to the "closed" crystallographic-like structure where there is little to no water present within 3 Å of the protonation site (**Figure 5.1.2.2a**). Using the V66K mutant as an example, the we demonstrate that SNase conformations sampled in our simulations interconvert between closed and open states (**Figure 5.1.2.2b,c**) when the external pH is close to the calculated pK_a value. While the coupling between pK_a and conformational sampling has been reported for a few isolated examples,[123,226] and the importance of sampling such alternative conformational states has been previously postulated by experiments,[228] there has been no comprehensive proof for their role. Our work presents the first interpretation that pH-dependent transient states not only exist, but may be of general importance for proteins with buried ionizable groups. For all mutants with highly shifted pK_a values, we observed that protonation of the internal Lys was concomitant with an increase in local solvation around the protonation site as illustrated in **Figure 5.1.2.2a**. The backbone RMSD of the entire structure between both closed and open states in these mutants is small, ranging from 1.2 to 1.9 Å, which suggests that the conformational relaxation to accommodate a buried charged residue does not require significant structural rearrangement. Our observations are consistent with experimental measurements that indicate buried ionizable residues in SNase are readily accommodated without any special structural adaptation or distortion to the overall protein conformation.[173,223]

5.1.3 Comparison to Implicit Solvent CPHMD Simulations

Next, we investigate the various computational models, specifically previous CPHMD implementations, to ascertain the robustness of predictions obtained from simulations. One insightful observation is that our calculated pK_a for the N100K mutant of SNase is 6.6, which is close to the calculated pK_a of 7.0 reported by Shen and co-workers,[123] despite the differences in the solvation model and the cutoff schemes utilized. This suggests that the CPHMD framework is not overly sensitive to specifics of the simulation setup. This is because CPHMD calculates the free energy of protonation relative to a reference compound, and, as long as the simulation of the protein and the reference are performed under identical conditions, the differences originating from the simulation setup approximately cancel out. To test this hypothesis, we compared our pK_a predictions from explicit solvent pH-REX CPHMD$^{MS\lambda D}$ simulations with those obtained from the GBSW[89] implicit solvent pH-REX CPHMD framework, using the refined protocol presented in this paper. Our results (**Fig. 5.1.3.1a,b** and **Table 5.1.3.1**) show excellent correlation of $R^2 = 0.89$ between the pK_a predicted from both explicit and implicit models. However, unlike the explicit solvent CPHMD$^{MS\lambda D}$ simulations, which resulted in a pK_a shift of more than 10 units when only the closed state was used, the implicit solvent CPHMD simulations produced pK_a shifts that were smaller in magnitude. The van der Waals surface representation used to define the solute-solvent dielectric boundary in GBSW is known to form small crevices of high dielectric region between atoms, which results in the underestimation of the Born radii and overestimation of the solvation free energy.[91] Therefore, despite the actual hydrophobicity of the environment near the protonating site, the GBSW model may be too "wet", causing a smaller pK_a shift. Alternatively, the same effect can be achieved as a result of the faster conformational dynamics in the GBSW model, which may sample both open and closed states more frequently.

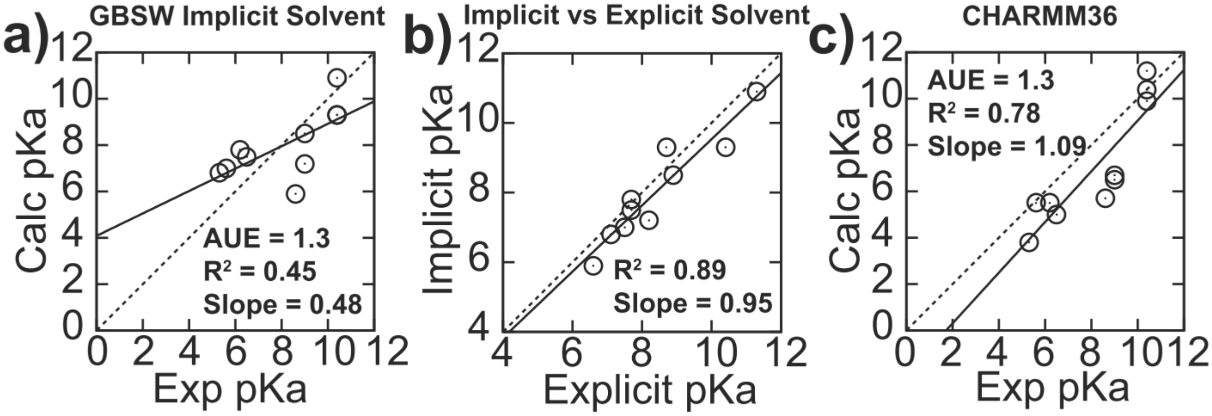


Figure 5.1.3.1: (a) Accuracy of calculated pK$_a$ from implicit solvent pH-REX CPHMD simulations are similar to explicit solvent results, and (b) both sets of calculated pK$_a$ values are highly correlated with each other. (c) Calculated pK$_a$ values from explicit solvent pH-REX CPHMD$^{MS\lambda D}$ simulations using the CHARMM36 force field results in improved correlation with experimental pK$_a$ values. All simulations were initiated using both closed and open structures

To distinguish between these two possibilities, additional simulations, with the structures of the V66K, V99K, L125K and I92K mutants rigidified by applying harmonic restraints to all heavy atoms, were performed, and the resulting pK$_a$ values were similar to those calculated without restraints (see **Table 5.1.3.1**). These results suggest that smaller pK$_a$ shift in the implicit solvent primarily stems from the "wetness" of the GBSW model, rather than its faster conformational dynamics. Lastly, we investigated the effect of using a recently released CHARMM36 all-atom force field for proteins on our pK$_a$ predictions, as it has been previously shown to yield superior reproduction of experimental dynamical data[131] We recalculated the biasing potentials for the common titrating residues of proteins using CHARMM36 force field, and revised pK$_a$ values, as shown **Figure 5.1.3.1c**, indicate an improvement in the predictive performance over the older CHARMM22/CMAP force field, with R^2 increasing from 0.52 to 0.78 and the slope of the regression moving from 0.52 to 1.09, while maintaining the same level of accuracy with an average unsigned error of 1.3 pK$_a$.

Table 5.1.3.1: Calculated pK$_a$ SNase Lys mutants obtained from GBSW implicit solvent pH-REX CPHMD simulations

Variant	**Exp pK$_a$**	**CHARMM22/CMAP (Closed Only & Rigid)**	**CHARMM22/CMAP (Closed Only)**	**CHARMM22/CMAP (Closed & Open)**
I92K	5.3	4.2	6.8	6.9
V66K	5.6	6.4	6.4	7.0
L125K	6.2	7.2	7.2	7.8
V99K	6.5	4.0	7.0	7.4
N100K	8.6	-	3.5	6.0
V39K	9.0	-	6.9	7.3
Y91K	9.0	-	7.5	8.5
A58K	10.4	-	8.9	9.1
N118K	10.4	-	5.8	9.4
A132K	10.4	-	11.0	11.0
AUE			**1.8**	**1.3**
R^2			**0.13**	**0.45**
Slope			**0.34**	**0.48**

5.1.4 Relevance of Transient States at Physiological pH

Having established that buried ionizable residues can trigger pH-dependent structural fluctuations, we extend our analysis to determining the relevance of these alternative states at pH 7. Using a two-state model that assumes a conversion between one dominant open and one dominant closed state, we can derive an equation describing the ratio of open to closed states (R$_{OC}$). The equilibrium constant of the deprotonation reaction for the open (K$_{open}$) and closed (K$_{open}$) states are given as

$$K_{open} = \frac{[\mathrm{Open}][\mathrm{H}^+]}{[\mathrm{OpenH}]} \tag{5.1.4.1}$$

$$K_{closed} = \frac{[\mathrm{Closed}][\mathrm{H}^+]}{[\mathrm{ClosedH}]} \tag{5.1.4.2}$$

Each equation can be rearranged to the following forms:

$$\frac{K_{open}}{[\mathrm{H}^+]} + 1 = \frac{[\mathrm{Open}] + [\mathrm{OpenH}]}{[\mathrm{OpenH}]} \tag{5.1.4.3}$$

$$\frac{K_{closed}}{\left[\mathrm{H}^+\right]}+1=\frac{\left[\mathrm{Closed}\right]+\left[\mathrm{ClosedH}\right]}{\left[\mathrm{ClosedH}\right]} \qquad (5.1.4.4)$$

Dividing equation **5.1.4.3** by **5.1.4.4** we obtained the following expression:

$$\frac{\left[\mathrm{Open}\right]+\left[\mathrm{OpenH}\right]}{\left[\mathrm{Closed}\right]+\left[\mathrm{ClosedH}\right]}\frac{\left[\mathrm{ClosedH}\right]}{\left[\mathrm{OpenH}\right]}=\frac{\frac{K_{open}}{\left[\mathrm{H}^+\right]}+1}{\frac{K_{closed}}{\left[\mathrm{H}^+\right]}+1} \qquad (5.1.4.5)$$

Which can be rearranged to obtain the ratio of open to closed states (R_{OC}), which is defined as the total concentration of open states relative to the total concentration of closed states:

$$\mathrm{R}_{\mathrm{oc}}=\frac{[\mathrm{Open}]+[\mathrm{OpenH}]}{[\mathrm{Closed}]+[\mathrm{ClosedH}]}=\frac{\frac{K_{open}}{[\mathrm{H}^+]}+1}{\frac{K_{closed}}{[\mathrm{H}^+]}+1}\frac{[\mathrm{OpenH}]}{[\mathrm{ClosedH}]} \qquad (5.1.4.6)$$

Equation 5.1.4.6 can be rewritten as a function of the pK_a of the open (pK_{open}) and closed (pK_{closed}) microscopic states.:

$$\mathrm{R}_{\mathrm{oc}}=\frac{\frac{10^{-pK_{open}}}{10^{-pH}}+1}{\frac{10^{-pK_{closed}}}{10^{-pH}}+1}\frac{[\mathrm{OpenH}]}{[\mathrm{ClosedH}]}=\frac{10^{-pK_{open}}+10^{-pH}}{10^{-pK_{closed}}+10^{-pH}}\frac{[\mathrm{OpenH}]}{[\mathrm{ClosedH}]} \qquad (5.1.4.7)$$

Note that the first term denotes the pH-dependent term of the equation (K^{pH} term), and the second term is pH-independent (K^0 term). It is also the ratio of the open to closed states where both states are in the protonated form. The K^0 term can be related to the free energy difference between the two states as such:

$$K^0=\frac{[\mathrm{OpenH}]}{[\mathrm{ClosedH}]}=\exp\left(-\frac{\Delta G}{k_B T}\right) \qquad (5.1.4.8)$$

Alternatively, it can also be expressed in terms of the various pK$_a$ values relevant to the system. From previous studies,[123] the microscopic pK$_a$ of both closed and open states are related to the apparent pK$_a$ (pK$_{app}$) of the system:

$$pK_{app} = -\log\left(\frac{[\mathrm{OpenH}]}{[\mathrm{OpenH}]+[\mathrm{ClosedH}]}K_{open} + \frac{[\mathrm{ClosedH}]}{[\mathrm{OpenH}]+[\mathrm{ClosedH}]}K_{closed}\right) \qquad (5.1.4.9)$$

From **equation 5.1.4.7**, a series of rearrangement leads to the expression [OpenH]/[ClosedH]:

$$10^{-pK_{app}}([\mathrm{OpenH}]+[\mathrm{ClosedH}]) = [\mathrm{OpenH}]K_{open} + [\mathrm{ClosedH}]K_{closed} \qquad (5.1.4.10)$$

$$[\mathrm{OpenH}]\left(10^{-pK_{app}} - K_{open}\right) = [\mathrm{ClosedH}]\left(K_{closed} - 10^{-pK_{app}}\right) \qquad (5.1.4.11)$$

$$\frac{[\mathrm{OpenH}]}{[\mathrm{ClosedH}]} = \frac{\left(K_{closed} - 10^{-pK_{app}}\right)}{\left(10^{-pK_{app}} - K_{open}\right)} = -\frac{\left(10^{-pK_{closed}} - 10^{-pK_{app}}\right)}{\left(10^{-pK_{open}} - 10^{-pK_{app}}\right)} \qquad (5.1.4.12)$$

Combining **equations 5.1.4.7** and **5.1.4.12** we derive the following expression, which is the final form of the R$_{OC}$:

$$\mathrm{R}_{\mathrm{OC}} = \frac{[\mathrm{Open}]+[\mathrm{OpenH}]}{[\mathrm{Closed}]+[\mathrm{ClosedH}]} = K_{pH} \times K^0$$
$$\text{where } K_{pH} = \frac{\left(10^{-pK_{open}} + 10^{-pH}\right)}{\left(10^{-pK_{closed}} + 10^{-pH}\right)}; K^0 = -\frac{\left(10^{-pK_{closed}} - 10^{-pK_{app}}\right)}{\left(10^{-pK_{open}} - 10^{-pK_{app}}\right)}$$

(5.1.4.13)

This function will be continuous across the entire pH range, under the conditions that the pK$_{app}$ lies between pK$_{open}$ and pK$_{closed}$. In the limit of pK$_{app}$ = pK$_{open}$ = pK$_{closed}$ the function is discontinuous, but a slight offset (i.e. -0.01 pK$_a$ units) can be used to model the effect of no or extreme pK$_a$ shift. This function can be decomposed into a pH-dependent (K$_{pH}$) and pH-independent (K^0) terms. The K$_{pH}$ term depends on the microscopic pK$_a$ of each state (pK$_{closed}$, pK$_{open}$) and external pH. The K^0 term can be physically related to the free energy difference of the open and closed states in their protonated form (see **equation 5.1.4.8**). However, as each form favors opposing protonation states, this free energy is usually not measured by experiments. Therefore, it may be advantageous for K^0 to be expressed as a function of the system's macroscopic or apparent pK_a (pK_{app}), which can be readily measured, and the microscopic pK_a of each state (pK_{closed}, pK_{open}). Based on the R_{OC} equation, one may also derive the fraction of the open state (F_{open}), which is a more intuitive metric for open states at a specific pH.

$$F_{\text{open}} = \mathrm{R}_{\text{OC}} / (\mathrm{R}_{\text{OC}} + 1)$$

Due to the rapid dynamics and/or low population of the minor state, it is often beyond the detection limits of experiments to establish the pK_a values of each microscopic state.[228] However, because the open state is solvent exposed, pK_{open} may be approximated as the reference pK_a of the free amino acid. Using the range of pK_a shift of 10-20 units recorded from spectrophotometric measurements of organic acids and bases in a low dielectric solvent,[219,220] we have conservatively assigned pK_{closed} to be shifted by 10. As shown in **Figure 5.1.4.1**, one can use the R_{OC} equation to derive the pH-dependent fraction of the open state for a series of hypothetical pK_{app} for buried Lys or Glu. Our analysis indicates that pH-dependent transient open states may contribute as much as 2% of the total population at pH 7 when the apparent pK_a of Lys is shifted by as much as 5 units, which appears to be the upper limit of pK_a shift as recorded in current literature.[173,223] For Asp and Glu, an apparent pK_a shifted by 5 pK_a units represents a 1% contribution of the transient state at pH 7. Thus, for the residues with highly shifted pK_a values, the low population transient states are likely to contribute significantly to the apparent pK_a, and, thus, need to be elucidated to correctly compute the apparent pK_a.

Although the existence of transient states involving buried ionizable groups does not necessarily imply a functional relevance, there is increasing precedence that the inclusion of transient states is needed to fully account for biological properties.[45-47] In the context of our work, we suggest that effect of such pH-dependent transient states will be pronounced when an ionizable group transitions between hydrophilic and hydrophobic environments, such as in membrane fusion processes, where activated/transient states have been postulated to play a crucial role.[48] In addition, traditional studies of catalytic mechanisms have always assumed that crystallographic structures correlate with measured pK_a, but, as we have shown, that may not be true for buried residues with highly shifted pK_a values. Moreover, the coupled relationship of both open and close states and their role in recapitulating macroscopic experimental observables suggest that structural analysis of buried residues should be performed from the perspective of looking at structural pairs, as opposed to the conventional approach of a single static ground state conformation. For such analyses, the equations we have provided will prove useful for a quick "back of the envelope" estimation of the population of proposed transient states. For example, one could use the experimentally measured apparent pK_{app} value to select the appropriate curve in **Figure 5.1.4.1**, and use it to estimate the fraction of the open state at a given pH, as a means to evaluate the plausibility of experimental characterization within the detection limits of the methods employed. Alternatively, it can also be used to estimate the pH range where the population of proposed transient states will enter the detection range of experiments.

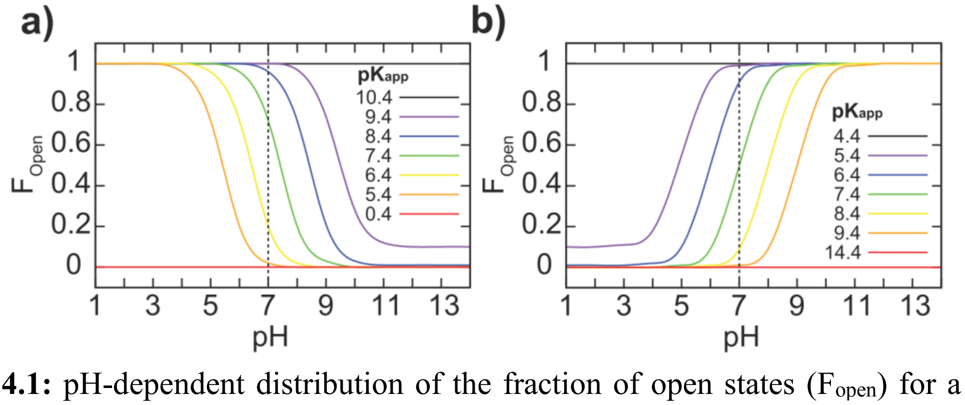


Figure 5.1.4.1: pH-dependent distribution of the fraction of open states (F_{open}) for a buried (a) Lys and (b) Glu residue. Several color-coded hypothetical apparent pK_a (pK_{app}) values are illustrated, shifted by 1 to 5 pK_a units. In the limits of extreme (i.e. 10) or null pK_a shift, the expected population of 100% closed or open states is recovered. Here, pK_{open} is 10.4 and 4.4, and pK_{closed} is 0.4 and 14.4, for Lys and Glu, respectively.

5.1.5 Conclusion: CPHMD$^{MS\lambda D}$ Simulations Identified pH-mediated Transient States in all Buried Ionizable Protein Residues

In this chapter, we used CPHMD$^{MS\lambda D}$ simulations to model the pH-dependent dynamics of a comprehensive set of SNase mutants with buried ionizable residues that have varying degrees of pK_a shifts. Among our key findings is that a buried charged residue cannot be accommodated inside a purely hydrophobic pocket and that an open state structure for these "buried" residues, characterized by local solvation around the protonating site, was observed in all SNase mutants with highly shifted pK_a. At physiological pH, buried ionizable groups with large pK_a shifts have transiently populated open states, where they contribute a small but non-zero population of 1-2% at pH 7. Nevertheless, sampling these open states is a necessary condition for accurately reproducing experimental pK_a measurements, to which calculated pK_a from our explicit solvent CPHMD$^{MS\lambda D}$ simulations demonstrated good agreement, with a low average unsigned error of 1.3 pK_a units and correlation coefficient of R^2 0.78. The work we present here provides the first validation that buried ionizable residues can readily trigger pH-mediated conformational fluctuations that may be observed as transient state structures at physiological pH. Lastly, the discovery of a coupled relationship of both open and closed states and their role in recapitulating macroscopic experimental observables suggests that structural analysis of buried residues may benefit from the perspective of looking at structural pairs, as opposed to the conventional approach of a single static ground state conformation.

5.2 The Role of Transient States in Tuning pH-Dependent Optical Properties of Cyan Fluorescent Protein

Note: Chapter 5.2 was adapted from the following references.[120] *The entire chapter 5.2 contains significant contributions from Elena Laricheva, who was responsible for performing the majority of the simulations. The results and discussion have been included in this dissertation for continuity and completeness.*

5.2.1 WasCFP: A Fluorescent Protein with an unusual pH-dependent Spectrum

Expanding the palette of genetically encodable fluorescent proteins (FPs) with spectral properties that can be modulated by pH is a well-appreciated challenge due to their wide applicability as non-invasive pH sensors[1–5] and optical highlighters for super-resolution imaging of living cells.[6–9] The majority of such proteins developed to date belong to the green fluorescent protein (GFP) family and owe their pH-sensitive optical behavior to a tyrosine-based chromophore that can interconvert between the neutral (protonated) and deprotonated (charged) states depending on the hydrogen-bonding environment surrounding its phenolic group.[7] Rational design of new pH-sensitive variants requires both (i) a fundamental understanding of how the proteins with tyrosine-based chromophores function at the atomic level, as well as (ii) going beyond and looking at the FPs with chromophores other than tyrosine as potential candidates (e.g. tryptophan or phenylalanine/histidine-based chromophores, as in the case of cyan and blue fluorescent proteins). While a second approach has long been overlooked, the first one has been quite successful resulting in a number of useful pH sensors (e.g. pHluorins,[3,5] phRed[2]) and optical highlighters (e.g. Kaede[8,9]). The efforts in this direction, however, have mostly been focused on targeting the residues in the vicinity of the chromophore that affect its spectral characteristics through electronic effects, and largely neglected the importance of characterizing the conformational ensemble of the protein.[7]

In recent years, a large body of evidence has emerged suggesting that understanding the mechanisms underlying protein functions depends on our ability to characterize its dynamic ensemble.[10–12] Due to the nature of conventional biophysical techniques that primarily probe the most stable protein conformers, our understanding has long been limited to the information regarding highly populated ground conformational states. However, such states often represent only one of the functional forms, and higher-energy physiologically-relevant conformers can be transiently populated (~10% or less) when initiated by external stimuli, such as substrate binding, pH changes, or thermal fluctuations.[12,13] While low-energy ground-state conformers residing at the bottom of the conformational energy landscape are normally separated by very small kinetic barriers and interconvert between one another within pico- to nanoseconds, the barriers between them and higher energy structures are larger and associated with micro- to millisecond timescale or longer. Recent advances in relaxation dispersion NMR spectroscopy[11,14] and room temperature X-ray crystallography[15] have made the detection of such transient conformational states possible, demonstrating their ubiquitous role in enzyme catalysis,[10] protein folding,[13,14] and ligand binding.[16] Transiently populated conformational states triggered by pH are of particular importance since pH regulates the biological activity of many proteins, and the role of pH-dependent transient states is an emerging discovery that has been recently reported to influence membrane fusion,[17] folding pathways,[18] and, more generally, the dynamics of buried ionizable groups.[19]

In this chapter, we demonstrate that pH-dependent transient conformational states can tune the absorption profile of cyan fluorescent protein (CFP)[20] – a blue-shifted variant of the green fluorescent protein (GFP) family used for multicolor labeling and fluorescence resonance energy transfer (FRET) applications – which to our knowledge is the first precedent of such a mechanism. In particular, we provide a theoretical explanation for the non-monotonic pH-dependent absorption of a recently engineered CFP mutant (WasCFP; see **Figure 5.2.1.1a**). While the vast majority of the pH-sensitive fluorescent proteins reported to date[2,3,5] exhibit monotonic changes in optical signals (e.g. absorption, emission, excitation), the WasCFP mutant does not conform to such a monotonic behavior. It reversibly interconverts between cyan-emitting and green-emitting forms, with the latter form dominant at pH 8.1 (and 25°C), above which the green signal drops.[21] Part of this behavior has been attributed to the deprotonation of the tryptophan-based chromophore (**Figure 5.2.1.1b**) at mildly basic pH, which is accompanied by a 60 nm bathochromic shift in absorption.

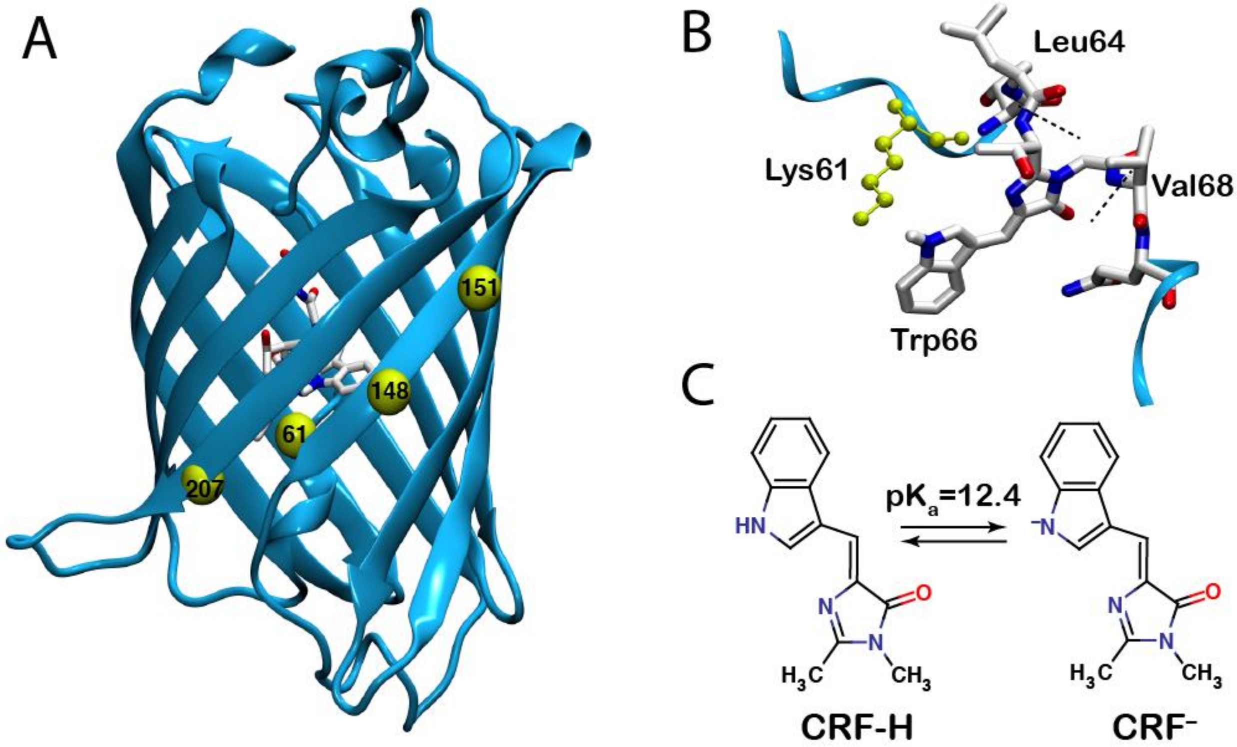


Figure 5.2.1.1: A: Structure of WasCFP showing C_α positions of mutated residues (V61K, D148G, Y151N, L207Q). B: Structure of Trp66-based WasCFP chromophore covalently bound to β-barrel at positions showed with dashed lines (Leu64 and Val68). C: Deprotonation of CRF. CRF-H and CRF$^-$ are neutral (protonated) and charged (deprotonated) forms of the synthetic chromophore, respectively.

Even though the observed effect is a highly unusual scenario as the pK_a of an analogous indole is high (16.2 in H_2O; 21.0 in DMSO), Sarkisyan *et al.*[21] have shown that a synthetic CFP chromophore (hereon referred to as CRF), which is a truncated version of that in the protein, undergoes a similar pH-dependent absorption red-shift in both protic and aprotic solvents, and its pK_a is depressed to 12.4 due to the more efficient delocalization of the negative charge over the extended π-system (**Figure 5.2.1.1c**). The same shift has been observed in a wild-type CFP variant mCerulean denatured in 5M NaOH. In addition, analogous pH-dependent bathochromic shifts, attributed to the deprotonation at phenolic oxygen, have been previously detected in various members of the GFP family with tyrosine-based chromophores (e.g. yellow, red and green fluorescent proteins).[22] In WasCFP, a key V61K substitution positions a Lys residue in close proximity to the indole nitrogen of the chromophore (**Figure 5.2.1.1b**), and this was proposed to stabilize its deprotonated state. The pK_a of the WasCFP chromophore, however, could not be directly measured and the atomic level details of its pH-dependent absorption remained unknown. Moreover, no explanation for the non-monotonic optical properties of WasCFP, specifically the signal drop in the green fluorescent form above pH 8.1 has been proposed.

5.2.2 Mapping out the Conformational States of WasCFP

First, we constructed a model compound (RES; denoting the model compound “residue”), which is an extended CFP chromophore consisting of the CRF moiety covalently bound to Leu64 and Val68 (see **Figure 5.2.1.1b** and **Figure 5.2.2.1**) to serve as a reference for our explicit solvent $CPHMD^{MS\lambda D}$ simulations.[25] Using a model pK_a of 12.7 (calculated using thermodynamic integration based on the CRF), we initially computed the pK_a values of the model wild type and mutant peptides (WTP and V61KP, respectively) that consist of 10 residues including a key position 61, before proceeding to the wild type and mutant proteins (WT and WAS, respectively) – all using a thermodynamic cycle depicted in **Figure 5.2.2.1**.

As shown in **Figure 5.2.2.1**, while an alchemical transformation of CRF to RES barely alters the pK_a of the titrating moiety, perturbation by the peptide and protein environment (WTP and WT) shifts its pK_a by 1.1 and 39.4 units, respectively.

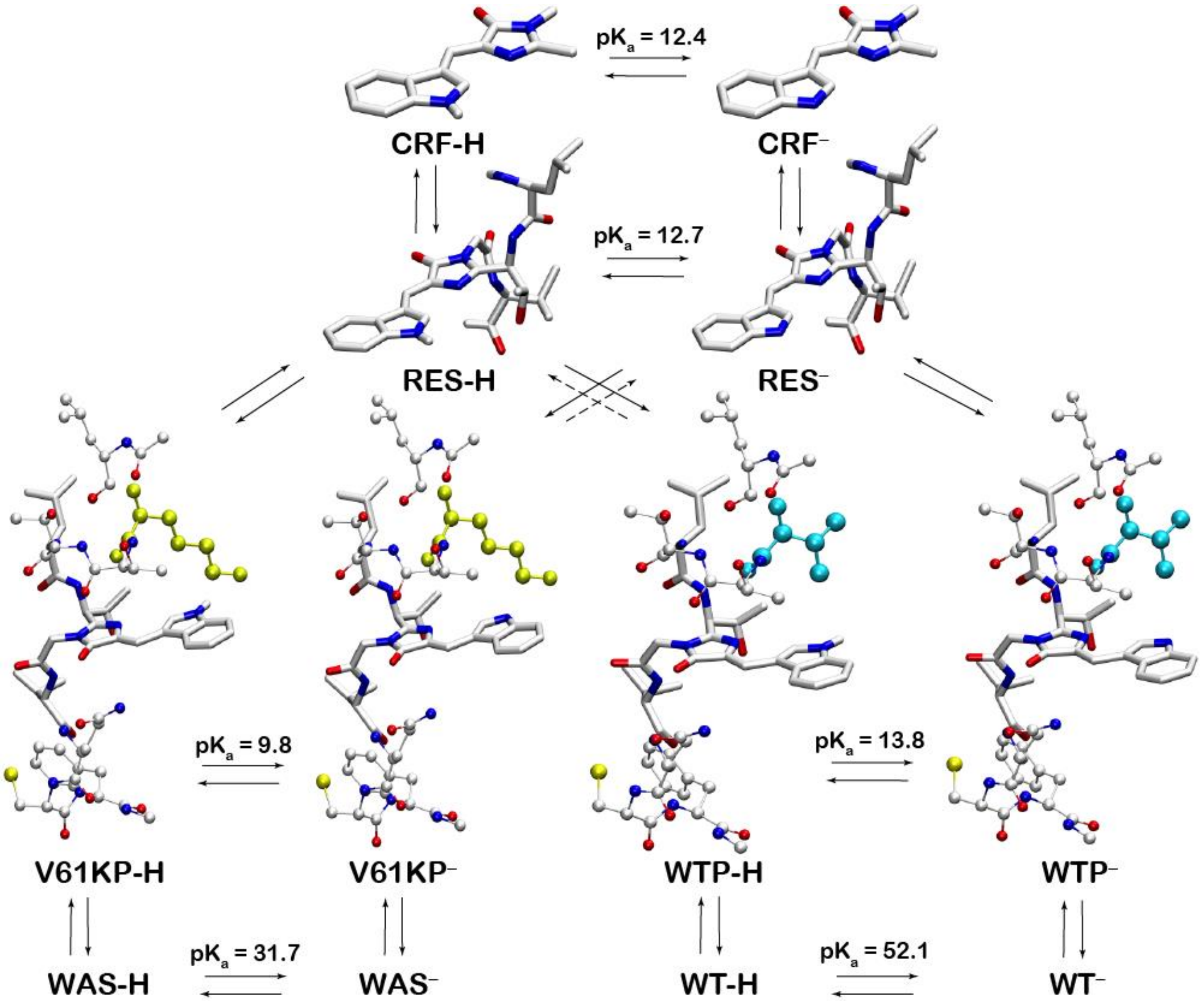


Figure 5.2.2.1. Thermodynamic cycle that shows alchemical transformations considered in this study. Cyan and yellow-colored side chains correspond to V61 and K61 in WTP and V61KP peptides, respectively. **pK_a** values computed from CPHMD$^{MS\lambda D}$ simulations are highly elevated in both WT and WAS and, thus, are not responsible the observed pH-dependent absorption of WasCFP.

Such large pK_a shifts have been reported when ionizable groups are transferred into a hydrophobic environment,[32] and demonstrates the sensitivity of the chromophore pK_a to the extent of local solvation at the protonation site. In our case, the solvation is high in the peptide and low inside the hydrophobic β-barrel of WAS, which is not surprising considering that nature selected the cylindrical β-barrel for fluorescent proteins in order to prevent the fluorescence quenching by either water or oxygen.[7] Using the thermodynamic cycle in **Figure 5.2.2.1** and the pK_a values computed using the CPHMD$^{MS\lambda D}$ method, we calculated that the positive charge of Lys61 introduced in a close proximity to the indole group of RES stabilizes V61KP by 5.4 kcal/mol with respect to WTP, and WAS – by 27.9 kcal/mol relative to WT. This cost is also solvation-dependent and leads to the downshift of the RES pK_a in a peptide by 4 units, while depressing the pK_a in the protein by 20.4. Even though our simulations clearly show that positive charge in the vicinity of the protonation site stabilizes the anionic form of RES, both neutral WT and charged WAS species are extremely stable, with pK_as of 52.1 and 31.7 that are comparable to those of the so-called "super-bases" in a low dielectric environment.[33] Therefore, the deprotonation at the indole nitrogen does not happen in such "closed" state conformations in the experimental pH range of 6–10, and the computed pK_a values do not account for the observed pH-dependent absorption properties of WAS.

Numerous studies of conformational plasticity of proteins, however, have demonstrated the importance of characterizing the dynamic ensemble of their states, including those that are only transiently populated.[10–13,16] Moreover, partially open, solvated pH-dependent transient states have been hypothesized to be of general importance in systems with buried ionizable groups.[19] However, capturing the transition between ground and transient conformational states often requires simulation timescales currently not accessible to the version of the CPHMD used in our work. Therefore, guided by the observation of a significantly diminished pK_a in a well-solvated peptide vs. “dry” high-pK_a closed conformation of the protein, we chose a hydration of the chromophore as our reaction coordinate and performed a search for an alternative WAS conformation using the weighted-ensemble sampling method[23] that allows the escape from deep local minima (high probability regions) and provides enhanced sampling of low probability, transient states. **Figure 5.2.2.2** shows a sampling of WAS configuration space along a hydration parameter (φ), which posits the existence of transiently populated states, characterized by the partially open β-barrel and local solvation at the protonation site. The pK_a values of the RES chromophore were subsequently computed for four representative WAS structures, extracted from the four regions in the histogram (φ=1.5-2.5; 5.5-8.0; 12.0-13.0 and 14.5-15.0), and structural changes influencing its pK_a were analyzed.

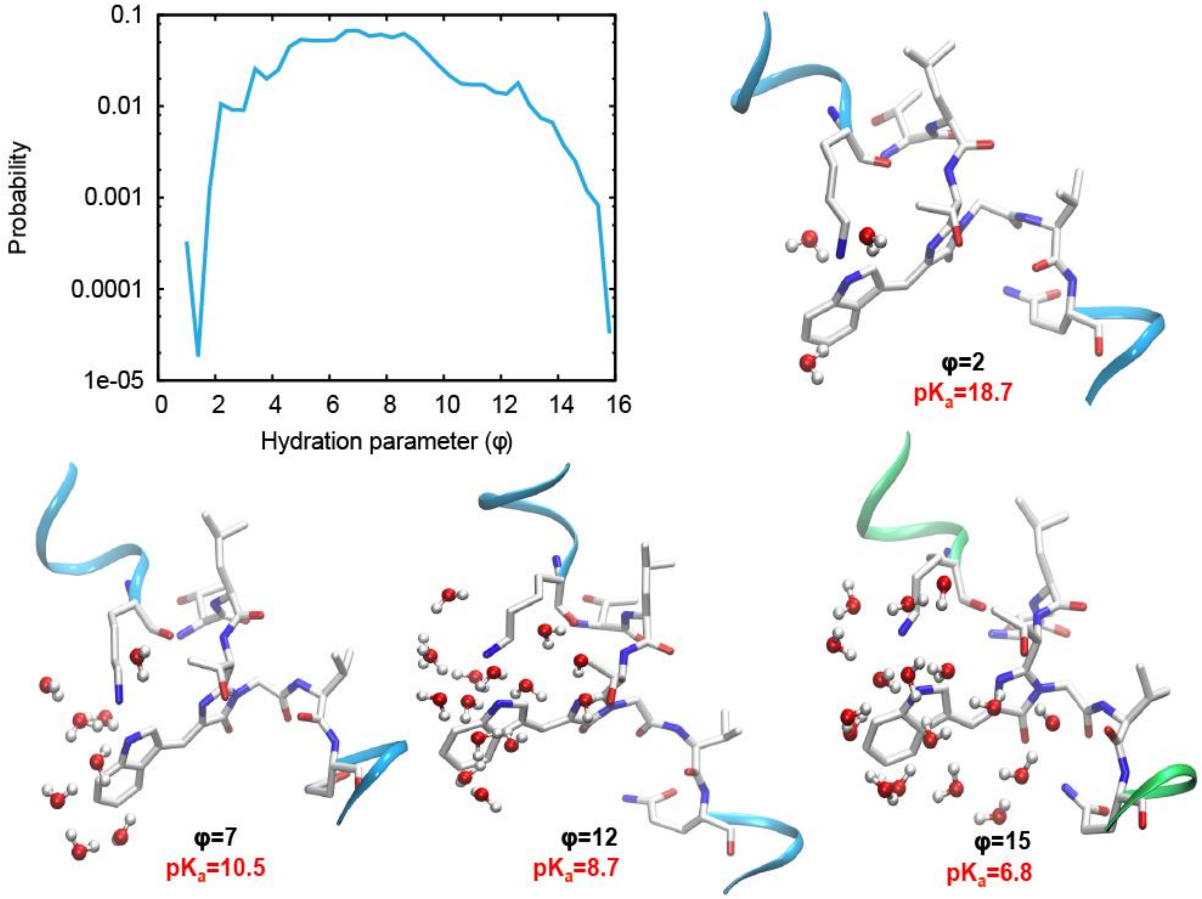
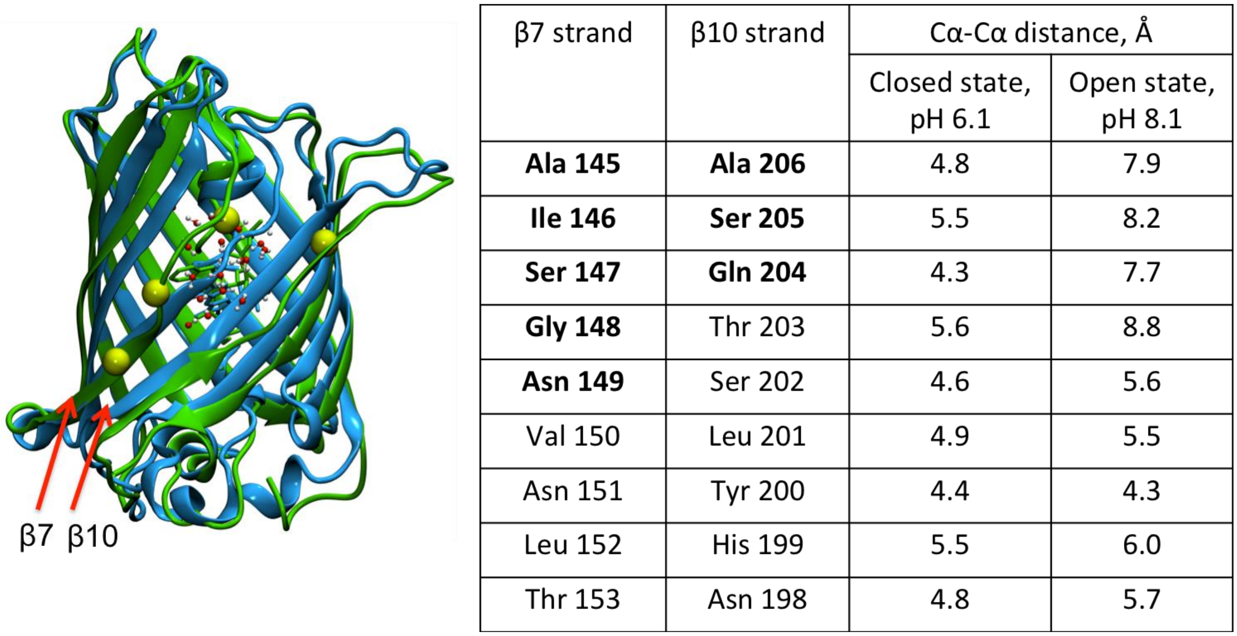


Figure 5.2.2.2: A. Probability distribution of the hydration parameter of RES in the WAS protein shows transiently populated states with large hydration parameters. B: Snapshots of the chromophore environment in four different conformations of WAS corresponding to hydration parameters φ=2, 7, 12, and 15. Number of water molecules within 7Å of nitrogen of the chromophore (3, 8, 15, and 17, respectively) and the corresponding pK_a values computed using CPHMD$^{MS\lambda D}$ simulations are shown for each conformation.



β7 strand	β10 strand	Cα-Cα distance, Å	
		Closed state, pH 6.1	Open state, pH 8.1
Ala 145	**Ala 206**	4.8	7.9
Ile 146	**Ser 205**	5.5	8.2
Ser 147	**Gln 204**	4.3	7.7
Gly 148	Thr 203	5.6	8.8
Asn 149	Ser 202	4.6	5.6
Val 150	Leu 201	4.9	5.5
Asn 151	Tyr 200	4.4	4.3
Leu 152	His 199	5.5	6.0
Thr 153	Asn 198	4.8	5.7

Figure 5.2.2.3. Distances between C_α-C_α atoms of residues in β7 and β10 strands in the open state (dominant at pH=8.1) vs. the closed state (dominant at pH=6.1).

We discovered that opening of the β7-β10 channel (up to 1.58Å backbone RMSD with respect to the crystal structure of WT; see **Figure 5.2.2.3**), which has been previously shown to be rather flexible in both wild type CFP[34] and in the study of the oxygen diffusion pathway in red fluorescent protein, mCherry,[35] facilitates local solvation (i.e. an increase in the number of water molecules within a 7 Å radius) at the indole nitrogen. Our calculations suggest that the pK_a of the chromophore can be as low as 6.8 for a structure with a high hydration parameter φ =15 that corresponds to as many as 17 water molecules within 7Å of the indole nitrogen of the chromophore (see **Figure 5.2.2.2**). Based on the structural data provided from our simulations, the residues on strands β7 and β10 undergo local unfolding into an unstructured loop. Notably, as illustrated in **Figure 5.2.2.3**, residues 145–149 of β7 and 204-206 of β10 lose their secondary structure. We note that these structural changes can be monitored by examining the carbon (C_α) chemical shifts in these regions as one titrates WasCFP from pH 8 to pH 6. We predict that these shifts would report a change from a mixed fraction of open and closed conformational states at pH 8 to a predominantly closed conformational state at pH 6. It is also worth noting that the chromophore in the open state is still significantly more rigid than it is in solution.

As WasCFP itself is a relatively recent construct, there is a lack of experimental data that measures its fluorescent properties as a function of its dynamics. However, parallels can be drawn between CFP and the related GFP, which, despite their differences in the chromophore, do share significant sequence similarity and may have similar conformational properties. Interestingly, prior studies of the unfolding of green fluorescent protein GFP have revealed a stable fluorescent intermediate that retained considerable secondary and tertiary geometry with displaced β7 and β10 strands and access of the water molecules to the chromophore,[36] It has also been noted that chromophore formation in fluorescent proteins occurs in a partially structured intermediate state, although this structure had reduced fluorescence. These experimental observations for GFP suggest that partially open conformations of the protein, similar to the transient state we observed in our simulations, can exist. In addition, in two companion papers, where the effect of pressure on the quenching of fluorescence was examined, Weber and co-workers[37] and Krylov and co-workers[38] reported that mCherry and mStrawberry are both highly fluorescent at standard pressure (0.1MPa) even though their chromophore are partially solvated. In fact, the extent of local solvation in our open state is similar to what the authors' computations predict for ambient conditions. In the context of our findings, it raises the possibility that non-ground state protein conformations may play a role in modulating spectral properties of fluorescent proteins.

5.2.3 The Role of Transient States in Modulating WasCFP pH-dependent Behavior

Lastly, to explain the unusual pH-dependent absorption behavior of WAS, we constructed a model based on the assumption that WAS interconverts between the hydrated transient (open) state that we identified and its original closed configuration. Previously, we developed a two-state model to explain the protonation equilibrium of SNase mutants with buried ionizable groups,[19] which allows one to compute the fraction of open state (F^{open}) as a function of pH based on the ratio of the open and closed state populations (R_{oc}) deduced from the simulations.

$$\mathrm{R}_{\mathrm{OC}} = \frac{[\mathrm{Open}]+[\mathrm{OpenH}]}{[\mathrm{Closed}]+[\mathrm{ClosedH}]} = K_{pH} \times K^{0}$$

$$\text{where } K_{pH} = \frac{\left(10^{-pK_{open}}+10^{-pH}\right)}{\left(10^{-pK_{closed}}+10^{-pH}\right)}; K^{0} = -\frac{\left(10^{-pK_{closed}}-10^{-pK_{app}}\right)}{\left(10^{-pK_{open}}-10^{-pK_{app}}\right)}$$

$$F_{\mathrm{open}} = \mathrm{R}_{\mathrm{OC}} / (\mathrm{R}_{\mathrm{OC}}+1)$$

This ratio is calculated using the computed microscopic pK_a values of both forms (pK_{open}=6.8 and pK_{closed}=31.7) and the apparent pK_a value estimated based on available experimental data (pK_{app}=7.8)

To test the validity of the model, we compared the computed F^{open} values with those that can be found directly from experiment. As a basis for our analysis, we used the pH-dependent absorption data for WasCFP recorded by Sarkisyan et al.[21] at 25°C (see Figures S2a and S2b of the Supplemental material in the original reference). The spectrum presented in their work differs from a typical pH-dependent absorption profile where one would expect for a mixture of conjugated acid and base, where one form is largely dominant at low pH, while the other one dominates under high pH conditions. In the case of WasCFP, the signal at 494 nm, corresponding to the charged (deprotonated) chromophore, grows with pH up to pH=8.1, but then its intensity decreases – due to titration of a nearby Lys61, as suggested by the authors of the paper. In addition, the signal of the protonated form consists of two peaks – a bonafide spectral feature of all cyan fluorescent proteins, the origin of which is still a subject of a great controversy.[20,29] While the authors do not mention any open states or fraction of the open states in their work and discuss the signals at 436 and 494 nm as arising from the protonated and deprotonated forms of the chromophore in the same protein conformation, our simulations suggest that chromophore can only exist in its deprotonated form when protein conformation is partially open allowing a few water molecules access to the deprotonation site. Therefore, for the remainder of this study, we will use the terms deprotonated and open states interchangeably and express the fraction of open state, F^{open}, using the following information only: (1) extinction coefficients at 436 and 494 nm ($\varepsilon_{436,\ HA}$ and $\varepsilon_{494,\ A-}$); (2) absorbance of the protonated (closed) form at the lowest pH, $Abs_{436,\ low\ pH}$ (neglecting a small absorption signal at 494 nm); and (3) absorbance of the deprotonated (open) form, Abs_{494} – which is the only variable in our final equation for F^{open}, presented below, that changes with pH.

Derivation of expression for F^{open} from experiment:

The concentrations of protonated (HA) and deprotonated (A^{-}) forms can be written in terms of the absorbances of the protein chromophore at the appropriate wavelengths (436 and 494 nm, respectively) using the Beer's-Lambert law:

$$Abs_{436} = \varepsilon_{436,\mathrm{HA}} \times b \times C_{HA} \tag{5.2.3.1}$$

$$Abs_{494} = \varepsilon_{494,A^{-}} \times b \times C_{A^{-}} \tag{5.2.3.2}$$

The sum of those concentrations represents the total concentration of all species, which remains constant at any pH:

$$C_{HA} + C_{A^-} = C_T \tag{5.2.3.3}$$

At low pH, the protonated form dominates, so that C_{HA} = $C_{T.}$ Similarly, at high pH there is predominantly deprotonated form, and C_{A^-}= C_T. Knowing that, we can express the absorbances at low pH and high pH as follows:

$$Abs_{436,\,\mathrm{low\,pH}} = \varepsilon_{436,\,\mathrm{HA}} \times b \times C_T \tag{5.2.3.4}$$

$$Abs_{494,\,\mathrm{high\,pH}} = \varepsilon_{494,\,\mathrm{A^-}} \times b \times C_T \tag{5.2.3.5}$$

From **equation 5.2.3.4** we obtain C_T and by substitution of **equation 5.2.3.3** into **equation 5.2.3.5**, the absorbance at high pH can be expressed in the following way:

$$Abs_{494,\,\mathrm{high\,pH}} = \frac{\varepsilon_{494,\,\mathrm{A^-}}}{\varepsilon_{436,\,\mathrm{HA}}} \times Abs_{436,\,\mathrm{low\,pH}} \tag{5.2.3.6}$$

By dividing **equation 5.2.3.2** by **equation 5.2.3.5,** we obtain the fraction of the deprotonated state, which varies with pH:

$$F^{open} = \frac{C_{\mathrm{A^-}}}{C_T} = \frac{Abs_{494}}{Abs_{494,\,\mathrm{high\,pH}}} = \frac{Abs_{494} \times \varepsilon_{436,\,\mathrm{HA}}}{Abs_{436,\,\mathrm{low\,pH}} \times \varepsilon_{494,\,\mathrm{A^-}}} \tag{5.2.3.7}$$

Both extinction coefficients are available from experiment:

$e_{436,\ \mathrm{HA}} = 28000\ M^{-1}cm^{-1}$ and $e_{494,\ \mathrm{A^-}} = 51000\ M^{-1}cm^{-1}$

The value of absorbance of pure HA at low pH, $Abs_{436,\ low\ pH}$, remains the same across the entire pH range, and the only parameter that varies is the absorbance of the A⁻ form at 494 nm (Abs_{494}), which can be calculated directly from the intensity of the peak. **Table 5.2.3.1** provides all the information necessary to estimate F^{open} at different pH using the information from the pH-dependent absorption spectrum at 25°C recorded by Sarkisyan et al.

Table 5.2.3.1. F^{open} derived from experimental data

	pH	Abs_{494}	F^{open}
$e_{436,\ HA} = 28000\ M^{-1}cm^{-1}$ $e_{494,\ A^-} = 51000\ M^{-1}cm^{-1}$ $Abs_{436,\ low\ pH} = 1.00$	6.1	0.16	0.09
	6.5	0.28	0.15
	7.0	0.54	0.30
	8.1	1.00	0.55
	8.5	0.91	0.50
	9.5	0.75	0.41
	9.9	0.45	0.25

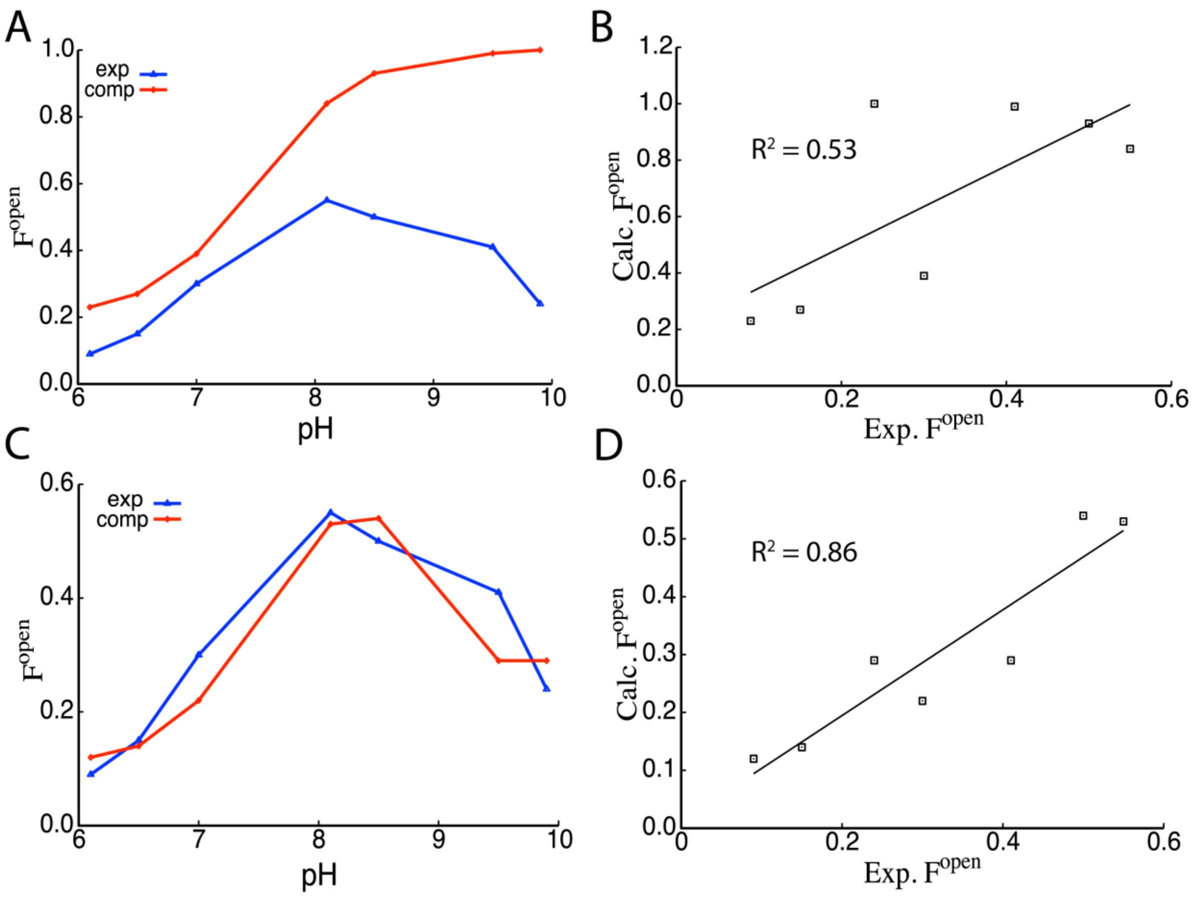


Figure 5.2.3.1. pH-dependent F^{open} computed using two (A) and three (C) state model. Correlation with F^{open} estimated from the pH-dependent absorption data is in (B) and (D), respectively.

As shown in **Figure 5.2.3.1b**, our two-state model provides a moderate correlation with experimental F^{open} values up to pH 8.1. However, it does not fully describe the system at higher pH (see **Figure 5.2.3.1a**), where Lys61 presumably deprotonates, as mentioned above.

Therefore, we introduced a third state that accounts for a neutral Lys61 and partially re-protonated chromophore at high pH, whose pK_a is approximated by that in the V61KP peptide found from our simulations (which is also in agreement with the experimentally suggested value of 9.8).

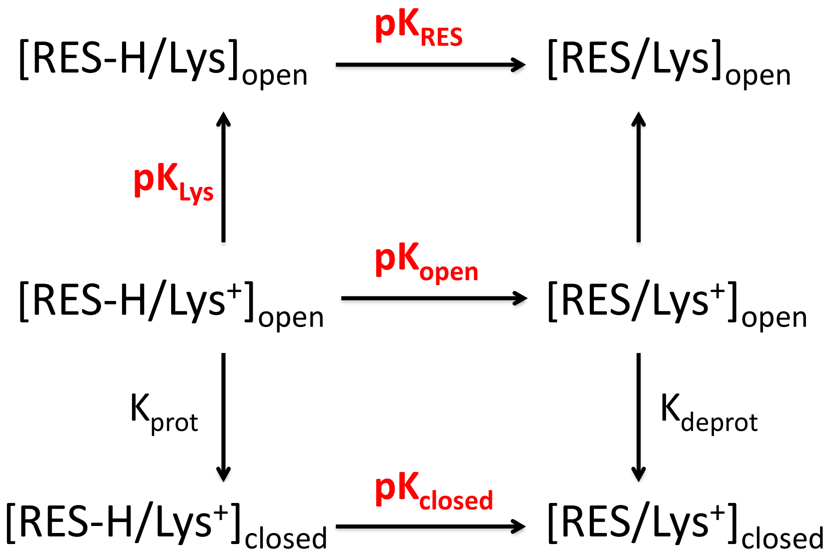


Figure 5.2.3.2. Schematic illustration of the three-state model. K_{prot} and K_{deprot} are equilibrium constants corresponding to the pH-independent conformational transitions between open and closed states.

The states in our three-state model are defined as follows: (i) $[RES\text{-}H/Lys^+]_{closed}$ – a closed state with neutral RES chromophore and protonated Lys61 (pK_{closed}=31.7); (ii) $[RES/Lys^+]_{open}$ – an open state with deprotonated chromophore and protonated Lys61 (pK_{open}=6.8); and (iii) $[RES\text{-}H/Lys]_{open}$ – an open state with re-protonated chromophore (pK_{RES}=9.8), and deprotonated Lys, whose pK_a (pK_{Lys}) was calculated from additional CPHMD$^{MS\lambda D}$ simulations where both the chromophore and Lys61 were simultaneously titrating. Due to the complexity of thermodynamic treatment of three different conformations, two of which have residues with strongly coupled protonation states (that of the chromophore and that of the nearby Lys61), we simplified the treatment of the three-state system into an effective two-state system. In our analysis, this effective two-state system corresponds to the bottom two arms of the thermodynamic cycle depicted in **Figure 5.2.3.2**, where the open state conformations are collectively represented by states (ii) and (iii) and where the protonation states of the chromophore and Lys61 are tightly coupled.

Using the microscopic pK_a calculated for states (i) and (ii), as well as the experimental apparent pK_{app} (7.8), we calculated that if the pK_a of the second state was shifted to 9.8 (which is the value representing state (iii) in the three-state model), it would result in a shift of pK_{app} to 10.8. Thus, the pK_{app} changes from 7.8 to 10.8 depending on the identity of the second state, which is primarily determined by the protonation state of Lys61. In other words, we assumed that the protonation state of Lys61 is coupled to the protonation state of the chromophore, which is justified because that is the only titrating residue in its vicinity. Hence, we interpolated the shift in the pK_{app} and pK_{open} by accounting for the fraction of the unprotonated Lys61. Thus, the corresponding pK_{app} and pK_{open} are expected to vary as a function of pH and we can substitute these pH-dependent terms to derive a ratio of the open to closed states for the three-state model. This ratio is then approximated using the modified R_{oc} equation:

$$R_{oc} = -\frac{(10^{-pK_{open}^{pH}} + 10^{-pH})(10^{-pK_{closed}} - 10^{-pK_{app}^{pH}})}{(10^{-pK_{closed}} + 10^{-pH})(10^{-pK_{open}^{pH}} - 10^{-pK_{app}^{pH}})}) \tag{5.2.3.8}$$

Where both pK^{pH}_{open} and pK^{pH}_{closed} are pH-dependent and differ from their corresponding values in the two-state model (pK_{open}=6.8 and pK_{closed}=31.7) by the delta term, used to interpolate the pK_a of the open conformation from state (ii) to state (iii) by accounting for the coupling between protonation states of the chromophore and Lys61.

$$pK_{open}^{pH} = pK_{open} + \Delta \tag{5.2.3.9}$$

$$pK_{app}^{pH} = pK_{app} + \Delta \tag{5.2.3.10}$$

$$\Delta = (pK_{RES} - pK_{open}) \times \frac{1}{1+10^{-(pH-pK_{Lys})}} \qquad (6.3.3.11)$$

As shown in **Figure 5.2.3.1d**, our three-state model not only improves the correlation with experimental observables (with R^2=0.86), but also captures, both qualitatively and quantitatively, the unusual bell-shaped pH-dependent absorption profile of WAS (see **Figure 5.2.3.1c**). Thus, our results show that the open state, collectively represented by the two locally solvated configurations and a partially open β-barrel, is transiently populated and contributes a small fraction at low pH (up to 12% at pH 6.1). This state becomes dominant (as much as 53%) at mildly basic conditions (pH=8.1), and gives rise to a strong absorption at 494 nm (and, thus, green fluorescence), which then titrates with a pK_a of 9.8 at higher pH values.

Lastly, our three-state model provides some useful insights into engineering pH-sensitive cyan fluorescent protein based on WasCFP. For example, to engineer a mutant that does not possess the residual cyan fluorescence at high pH, one may target the electrostatic environment in the vicinity of Lys61 in a manner to prevent its re-protonation at mildly basic conditions. Such a design would suppress the population of the third state in our three-state model by reducing the mechanism of the engineered mutant to two interconverting states, and this will increase the fraction of the open state with pK_{open}=6.8 (that only reaches a maximum of 53% at pH=8.1 in the current WasCFP design).

5.2.4 Conclusion: pH-mediated Transient States Identified from CPHMD$^{MS\lambda D}$ Simulations Account for the Non-Monotonic Optical Properties in WasCFP

In this chapter,, we have applied a combination of the weighted-ensemble sampling method[23] with a novel hydration parameter and explicit solvent CPHMD$^{MS\lambda D}$ simulations[25] to elucidate the origin of the unusual non-monotonic pH-dependent absorption behavior of a recently engineered CFP mutant that features a pH-dependent shift between cyan and green fluorescent forms. In earlier experimental observations, this optical property was proposed to be controlled by the charged anionic state of its tryptophan-containing chromophore.[21] Our calculations demonstrate that even in the presence of the stabilizing V61K mutation, the free energy cost of deprotonating the chromophore is still high and does not allow for the existence of the charged state of WasCFP even at basic pH. Instead, we propose the following explanation: The distribution between two transiently populated conformational states characterized by a partially open β-barrel with local solvation around the chromophore (that have pK_a values of 6.8 and 9.8), relative to the ground state crystallographic structure (that has pK_a of 31.7), is able to fully recapitulate both qualitatively and quantitatively the unusual non-monotonic pH-dependent properties of WasCFP. In this model, the open state is transiently populated at low pH, but reaches a population of 53% under mildly basic conditions, before losing its dominance and reverting to a transient state under highly basic conditions, and such mechanistic understanding may be used to further engineer the pH-sensitive fluorescent properties of WasCFP. Therefore, our work not only validates that tryptophan can be deprotonated in a biological system at mildly basic pH, but, more importantly, shows that pH-dependent transient conformational states are functionally relevant, and that they can tune the optical properties of fluorescent proteins. Such an outlook will have implications in the rational design of fluorescent proteins with pH-dependent optical properties.

Chapter 6: Considerations in a High Dielectric Environment

6.1 Parameter Validation

Note: Chapter 6.1 was adapted from the following references.[229]

6.1.1 Sodium Dodecyl Sulfate: A Model System for Charged Amphiphiles

As demonstrated in **Chapter 3.4**, developing accurate parameters for polyionic systems such as nucleic acids can be challenging, due to the possible limitations of using a classical force field to describe the electrostatics of such high charge density environments. This challenge may also be encountered in similar systems, such as structures based off charged amphiphiles that include membrane/lipids and other chemically similar surfactant compounds. A substantial effort by Shen and co-workers have extended the existing CPHMD framework to model the pH-dependent structural changes of surfactant-type systems (e.g. fatty acids). Such systems are increasingly used in the emerging field of nanomedicine in the development of nanocarrier devices to deliver drugs, genes, or other chemicals of interest, to specific malignant cells. Notably in cancer treatment, one can take advantage of the distinct pH profile of tumors to develop pH-sensitive nanocarrier devices that can release encapsulated drugs or other therapeutic agents by undergoing pH-dependent destabilization of the liposomal membrane.

Sodium dodecyl sulfate (SDS) is perhaps the most widely studied anionic surfactant, and it undergoes a number of structural changes as a function of its environmental factors, such as concentration and pH. For example, the structure of SDS-based solutions is dependent on its concentration. At the first critical micelle concentration (CMC) at 0.008M, it forms spherical/ellipsoidal micellar structures, which aggregate into larger rod-like micelles at its second CMC (~0.069M).[230-232] Eventually, at concentrations above 1.25M, higher-order structures such as lamellar phases start to form. The rheology and structure of SDS-based micelles are sensitive to temperature, cosurfactants, counterions, and even pH.[230-232] Thus, the inclusion of different species of cosurfactants and counterions at varying concentration, temperature and/or pH presents an opportunity to modulate the underlying physical properties of SDS-based solutions.

In this chapter, we examine various ion models developed for non-polarizable force fields, with the goal of identifying best practices in parameter validation and selection to ensure the accuracy of MD simulations of ionic surfactants. Notably, recent studies by Tang et. al.[233] have demonstrated the validating the accuracy of parameters can be size dependent. For example, in the larger aggregates of ~400 SDS molecules, different force fields produced different SDS morphology, but such a phenomenon was not detectable in simulations of smaller SDS micelles of 60 to 100 molecules.[233] One of the key diagnostics that Tang identified was the importance of intermolecular interactions between ionic species and the solvent, where selecting an appropriate set of parameters could reproduce experimentally observed SDS micelle morphology.[233] However, such a process requires prior knowledge about the corresponding micelle structure, which may not be an optimal solution if one is using MD simulations to make predictions in the absence of experimental data. The sensitivity towards intermolecular parameters, specifically in the anomalous “crystalline patches” that were formed in larger SDS micelles are reminiscent of excessive ion pairing which is a known issue with existing force fields.[234,235] Therefore, we hypothesize that the observed artifacts are a consequence of poor and/or inaccurate modeling of ionic species at high concentrations.

Traditional ion models that are used in all previous simulations of SDS,[233,236-239] hereon referred to as 1st-gen models, were parameterized against experimental data representative of infinite-dilution conditions.[240-243] These parameters are known to cause excessive ion pairing at high concentration.[234,235] More recent developments have focused on using experimental measurements that are more representative of finite concentration environments, using Kirkwood-Buff Integrals (KBI)[244-246] or osmotic pressure.[247-249] In the Kirkwood-Buff inspired reparameterization of 1st-gen ion models, hereon referred to as GROMOSKBFF, an additional scaling factor had to be introduced to modulate the interactions between Na^+ and SPC/E water in order to reproduce the experimental KBI at high concentration (see **Table 6.1.1.1**).[244-246] Similarly, in the osmotic pressure reparameterization process, an additional pairwise interaction potential, also termed NBFIX (non-bonded fixes) in the CHARMM program[124] had to be introduced between specific ion pairs (see **Table 6.1.1.1**). In both the KBFF and NBFIX approaches, hereon referred to as 2nd-gen ion models, they share the similarity of introducing these "additive hacks" that effectively break the conventional combination rules used by the force field.

Table 6.1.1.1: Parameters used in GROMOSKBFF[244,245] and NBFIX corrections to the CHARMM force field.[247,249] For GROMOSKBFF, the following combination rules were used: $\sigma_{ij} = \sqrt{\sigma_{ii} x\, \sigma_{ij}}$, $\varepsilon_{ij} = \sqrt{s(\varepsilon_{ii} x\, \varepsilon_{ij})}$ where s is the scaling factor for interactions between cations and water (s = 0.75 for Na, s = 0.80 for K). For CHARMM force fields, the following combination rules were used: $R_{ij} = (r_i + r_j)/2$, $\varepsilon_{ij} = \sqrt{\varepsilon_i\, \varepsilon_j}$. For NBFIX corrections applied, the R_{ij} values of specific atom pairs (listed above) are overridden by specially parameterized R_{ij} values.

Model	**Atom**	**σ_{ii} (nm)**	**ε_{ii} (kJ/mol)**	**ε_{iO} (kJ/mol)**	**q (e)**
GROMOSKBFF	Na	0.1820	0.3200	0.3420	+1.0
	K	0.2450	0.1300	0.2327	+1.0
Model	**Atom Pair**	**R_{ij} (Å)**	**E_{ij} (kcal/mol)**		**q (e)**

NBFIX	Na-Cl	3.731	−0.08388		+1.0
	Na-O2L	3.160	−0.07502		+1.0

6.1.2 On the Importance of Local Concentration and Finite Concentration Target Data

While the issues of modeling ionic species at higher concentration is known,[234,235] even in the largest SDS construct tested, the bulk concentration of SDS is comparatively low at 0.26M.[233] We suggest that a shift from analyzing the global concentration of the system to the local concentration of ionic species at the interface between the surfactant and the solvent needs to be realized. To demonstrate this conceptually, we examine a bilayer and spherical micelle structure, which are representative of high and low SDS concentrations respectively. Using a hypothetical bilayer segment of surface area 30 Å x 30 Å, and considering the region of ±6 Å at its boundary, we calculated that the local volume around the surfactant-solvent interface is 10,800 $Å^3$. Using the van der Waals volume of a SDS head group (25 $Å^3$) as the upper limit of surfactant packing, and a more realistic surface area of 40 $Å^3$, the estimated local concentration ranges from 3.4M to 5.5M. For a spherical SDS micelle of aggregation number 60, the typical radius is 15 Å, and using a similar means of evaluating its local volume (36,000 $Å^3$), the estimated local concentration is 2.7M. Therefore, across the concentration range of SDS and the corresponding structures that are formed, the local concentration of SDS falls between 2M to 4M, which is much higher than the sub 1M environment in which most parameters have been validated.

Next, we investigate the behavior of sodium chloride, which is perhaps the most basic model that has been parameterized extensively in all major force fields. KBI values and osmotic pressure at varying concentrations were calculated for the following force fields: CHARMM36, CHARMM36 with NBFIX, GROMOS45A3, GROMOS53A6 and GROMOSKBFF. As shown in **Figure 6.1.2.1a-b**, the accuracy of 1st-generation ion parameters in reproducing KBI values of

NaCl start to degrade at concentrations over 1M. In contrast, both 2nd-generation ion parameters, CHARMM36 with NBFIX and GROMOSKBFF, produce KBI values that are in much better quantitative agreement to experiments. For the osmotic pressure calculations, the results from CHARMM36 with NBFIX is similar to that reported by Luo and Roux,[247] reaching a projected value of ~300 bar at 5M. A comparison of concentration-dependent osmotic pressure across the various force fields, as illustrated in **Figure 6.1.2.1c**, yielded a similar observation to that of the KBI results. It should also be noted that GROMOSKBFF was parameterized against a single data point (KBI at 4.0M), and CHARMM36 with NBFIX was parameterized against osmotic pressure measured at multiple concentration values. Thus, a substantial portion of the simulations we have performed are not part of the target data in the parameterization process of these 2nd-gen ion models. To determine the structural reasons behind the ability of various force fields to reproduce KBI values and osmotic pressure, we analyzed the RDF of NaCl. In the 1st-gen ion models, excessive ion pairing was observed (**Figure 6.1.2.1d**), as evident from the strong first peak and/or multiple secondary peaks that suggest long-range ordering of Na^+ and Cl^- ions. In comparison, both 2nd-gen ion models had RDF peaks with significantly reduced heights.

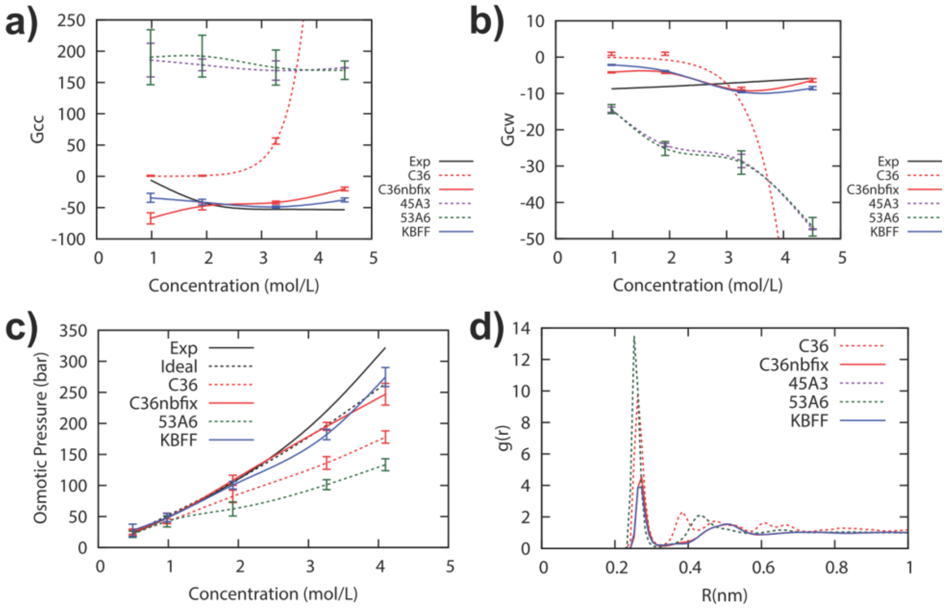


Figure 6.1.2.1: Concentration dependent Kirkwood-buff integrals (KBI) of NaCl for the (a) solute-solute, Gcc and (b) solute-solvent, Gcw terms, and (c) concentration dependent osmotic pressure of NaCl. Fitted lines were added to improve visualization, and error bars denote the standard deviation across 3 independent runs. All 1st-gen ion parameters (C36, 45A3, 53A6) failed to reproduce experimental KBI values, and 2nd-gen ion parameters (C36nbfix, KBFF) demonstrated a significant improvement in accuracy. (d) RDF describing the ion distribution of 4.50M NaCl indicates excessive ion-pairing in 1st-gen ion parameters.

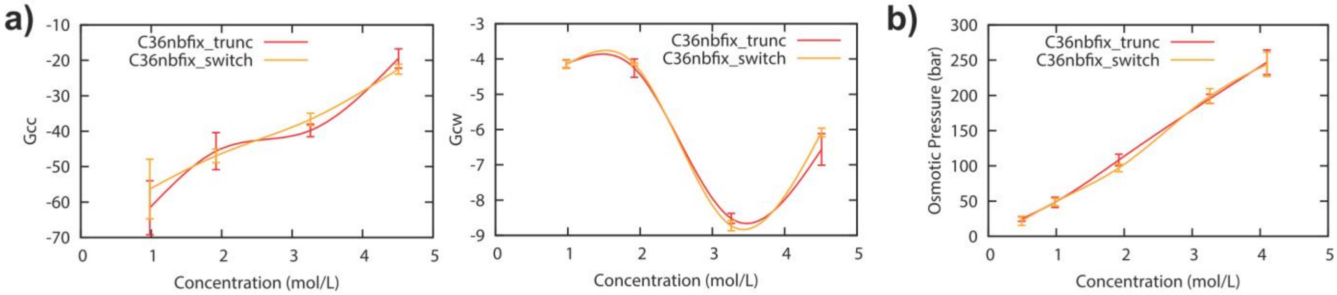


Figure 6.1.2.2: Concentration dependent (a) KBI values and (b) osmotic pressure of NaCl using different cutoff schemes, with error bars denoting the standard deviation across 3 independent runs.

Next, we investigated the effect of cutoff schemes on the calculated KBI and osmotic pressure. Traditionally, simulations performed in GROMACS using the GROMOS force field (including the development of GROMOSKBFF) use a shorter real space cutoff with Ewald sum and applies a simple truncation of VDW interactions. In comparison, CHARMM simulations (including the development of CHARMM36 with NBFIX) uses a switching function from 10-12 Å for VDW interactions, and Ewald sum for the treatment of long-range electrostatics. Our simulations have been performed using the former settings, and therefore our simulations using the CHARMM force fields may be performed under “incompatible” settings. Therefore, we recomputed the KBI values and osmotic pressure of NaCl using the CHARMM36 with NBFIX force field to determine if differences in the cutoff schemes used produce any discernable effect on the results. As shown in **Figure 6.1.2.2**, to within the precision of our simulation (across 3 independent runs), the results are identical, and this demonstrates that the ion parameters are not overly sensitive to the cutoff scheme differences employed in typical GROMACS and CHARMM simulations. Nevertheless, we caution that this insensitivity is only observed for the set of cutoffs that we have tested, and more drastic changes to the treatment of non-bonded interactions, such as changing from Ewald to Reaction Field for the treatment of long-range electrostatics, or changing from a Lennard-Jones 6-12 potential to a Buckingham potential for the VDW interactions could have more dramatic effects.

6.1.3 Parameter Validation in a Size Independent Manner

Methyl sulfate, which has an analogous functional group to the head group of SDS, serves as a model compound for SDS, and using truncated head groups as a model compound for parameter development has been reported previously.[250] To date, there has been no explicit effort to parameterize sulfate groups against experimental osmotic pressure or KBI. However, in the CHARMM36 force field, the ionic oxygens on sulfate groups share the same atom type as the ionic oxygens on phosphate groups, for which Pastor, Roux and co-workers have recently implemented NBFIX parameters for interactions between the phosphate ionic oxygens and Na^+ ions.[249] Thus, we used these NBFIX parameters for sodium methyl sulfate, and compared our computed osmotic pressure to experimental data.[251]

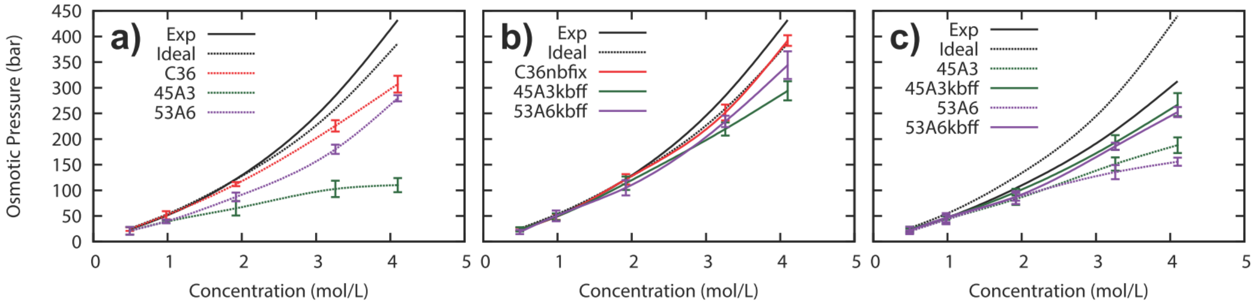


Figure 6.1.3.1: Concentration dependent osmotic pressure of sodium methyl sulfate using (a) 1st-generation and (b) adapted 2nd-generation ion parameters. All parameters with the exception of GROMOS45A3 demonstrated reasonable agreement with experiments. (c) Concentration-dependent osmotic pressure of potassium methyl sulfate.

As shown in **Figure 6.1.3.1a-b**, the default parameters from CHARMM36 are capable of reproducing experimental osmotic pressure reasonably well up to ~2M. This perhaps explains why the CHARMM36 SDS micelle simulations by Tang et. al. yielded reasonable agreement with experimental observation.[233] With the inclusion of the NBFIX parameters, it improved the agreement up to the simulated concentration of ~4M, reinforcing our earlier observations in NaCl simulations that NBFIX corrections are necessary to ensure accurate reproduction of osmotic pressure at higher concentration. For the GROMOS force field series, it was previously demonstrated that micelle structure was most significantly influenced by intermolecular parameters (van der Waals and partial charge parameters).[233] Therefore, for this study we created a modified GROMOS53A6 parameter set that uses the intermolecular parameters from GROMOS53A6 but retains the intramolecular parameters of GROMOS45A3. The results, as shown in **Figure 6.1.3.1a**, indicate that GROMOS45A3 significantly underpredicts the osmotic pressure of sodium methyl sulfate at concentrations above ~1M. In contrast, the osmotic pressure calculated from GROMOS53A6 is in much better agreement with experimental values. To determine if the existing SDS parameters in the GROMOS force field series can be improved, we combined the methyl sulfate parameters for both GROMOS45A3 and GROMOS53A6 with the ion parameters of the GROMOSKBFF force field, which we denote as GROMOS45A3KBFF and GROMOS53A6KBFF respectively. As shown in **Figure 6.1.3.1b**, the inclusion of KBFF modifications without any further adjustments significantly improved the agreement with experimental osmotic pressure, to the extent that our results suggest that a KBFF correction to the “inaccurate” GROMOS45A3 force field should yield similar performance to the force fields that produced the correct SDS morphology.

In order to attain a better understanding of the underlying reasons for the ability (or inability) of the various force fields to reproduce experimental osmotic pressure, we analyzed various RDFs that describe the solvent structure around sodium methyl sulfate (MESU). In the

Na^+-MESU(O) RDF, a strong first peak was observed in GROMOS45A3, which is indicative of excessive ion pairing, and this property is conspicuously absent in all other force fields tested. This observation correlates with the artifacts observed by Tang et. al., where only the GROMOS45A3 simulations formed crystalline lattice patches of SDS head groups and Na^+ ions.[233] When we examined the Na^+-SPC/E and MESU(O)-SPC/E RDFs, we observed that the strength of the hydration shell is unusually low for GROMOS45A3, and the 2nd-gen ion parameters had the strongest hydration shell behavior. This trend is similar to the analysis reported by Tang et. al. who tested various combinations of ion and water models and concluded that the models with a stronger first solvation shell would produce correct SDS micelle morphology.[233] However, unlike the prior work where the choice of ion and water models was systematically explored to find the best combination, our approach suggests that a model which produces good agreement with experimental osmotic pressure is likely to produce accurate SDS micelle structures, thus providing a physical principle and proper justification for the development and selection of parameters for simulating charged surfactant systems.

While we have demonstrated that 2nd-gen ion parameters are capable of reproducing the experimental properties of sodium methyl sulfate, it is also desirable to have a model that can accurately distinguish interactions between different cations. One example is the Hofmeister series, which is a classification of an ion’s ability to affect the solubility of proteins, presumably through their interactions with the side chain functional groups of the protein. Recent developments in KBFF-based force field reported by van der Vegt and co-workers has demonstrated that 2nd-gen ion parameters are able reproduce the Hofmeister series in the context of cation specific binding with protein surface charges.[246] In the context of our work, we simulated potassium methyl sulfate, which is a particularly challenging system as it exhibits an inverse behavior compared to sodium methyl sulfate, where potassium methyl sulfate has a higher tendency to form ion-pairs than sodium methyl sulfate, and consequently has a lower osmotic pressure than the ideal value at higher concentrations.[251] This is reflected by the fact that the experimental osmotic pressure of potassium methyl sulfate is lower than its ideal value at higher concentration, whereas that of sodium methyl sulfate is higher than its ideal osmotic pressure. **Figure 6.1.3.1c** illustrates the ability of various GROMOS force fields to reproduce the experimental osmotic pressure of potassium methyl sulfate. As before, the inclusion of the KBFF corrections improve the results, and reasonably good agreement with experimental osmotic pressure is achieved up to a concentration of ~3M. Our results on methyl sulfate suggest that the use of 2nd-gen ion parameters is not only accurate enough to model SDS surfactants in the presence of counter ions, it may also be sufficiently accurate to distinguish interactions based on the identity of the interacting monovalent ion.

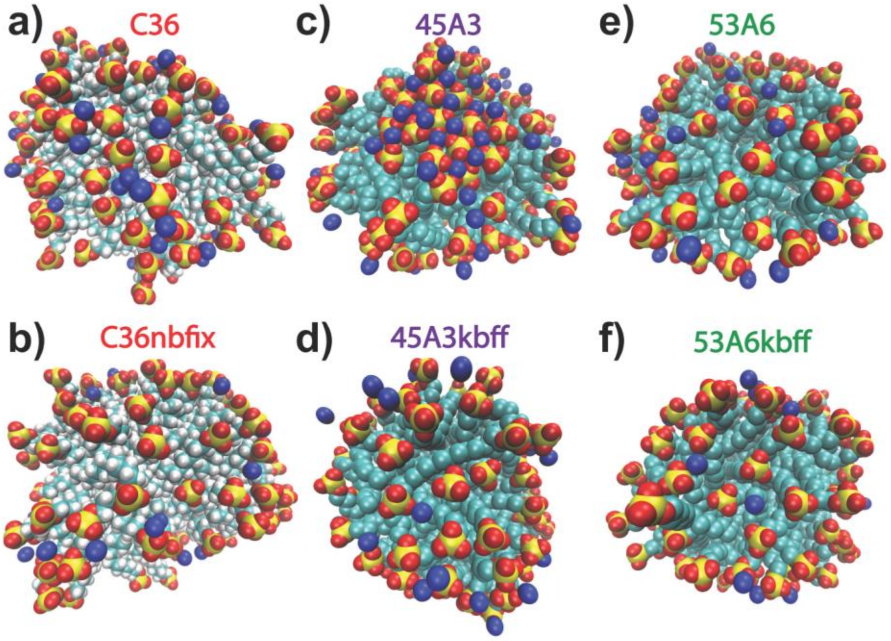


Figure 6.1.3.2: Representative snapshots of a SDS micelle with an aggregation number of 100 for (a) CHARMM36, (b) CHARMM36 with NBFIX, (c) GROMOS45A3, (d) GROMOS45A3KBFF, (e) GROMOS53A6 and (f) GROMOS53A6KBFF force field. With KBFF corrections applied to GROMOS45A3, the anomalous crystalline lattice patch formation was not observed.

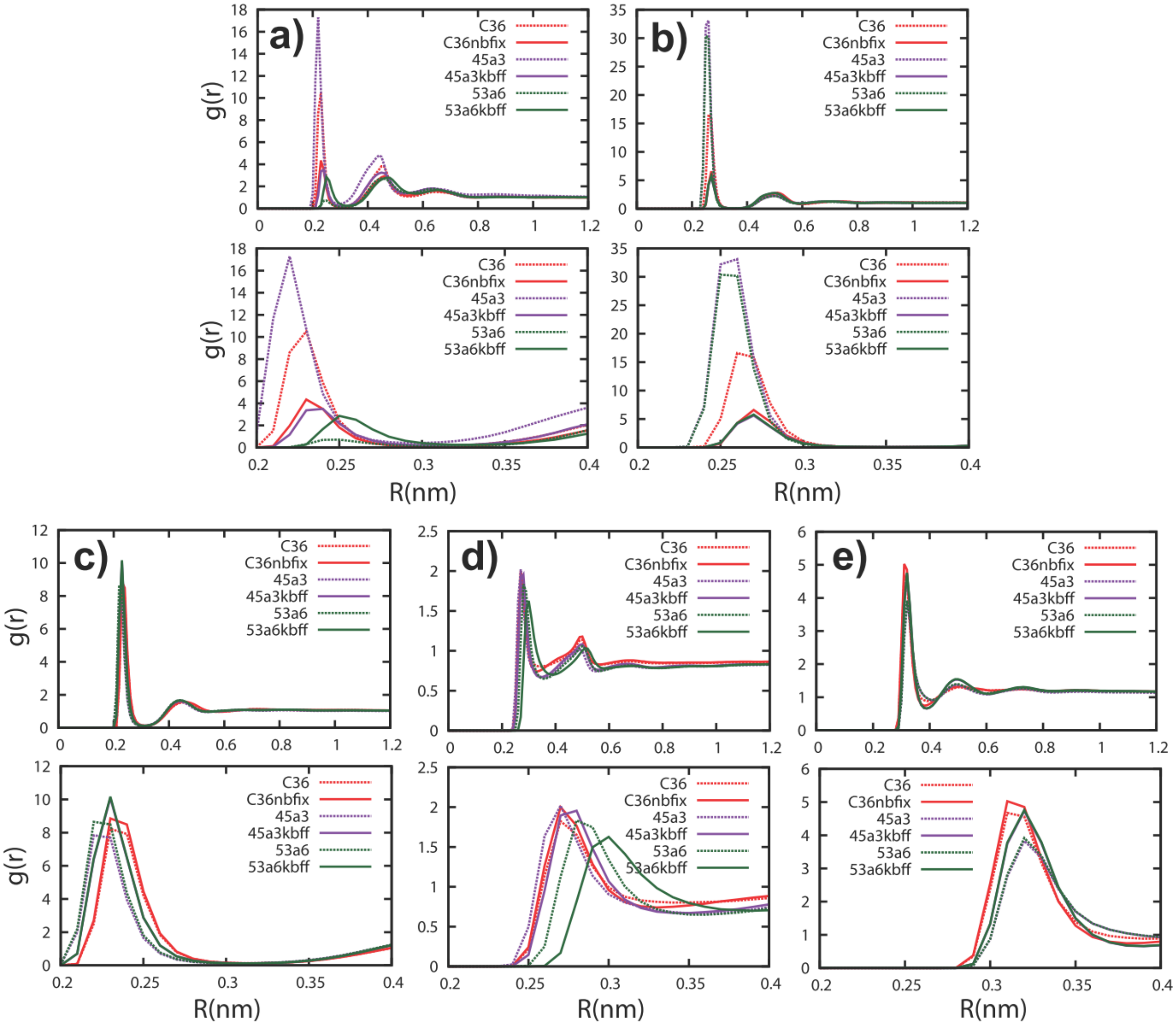


Figure 6.1.3.3: RDF of (a) Na^+ to SDS(O) and (b) Na^+ to Cl^- indicates that 1st gen ion models have a much higher propensity to form ion pairs than 2nd-gen ion models. RDF of (c) Na^+ to SPC/E, (d) SDS(O) to SPC/E and (e) Cl^- to SPC/E indicates a more subtle change in the solvation patterns

Having demonstrated that most of the force fields we tested, with the exception of GROMOS45A3, are able to reproduce the experimental osmotic pressure of sodium methyl sulfate reasonably well up to moderate concentrations of ~2M, we now focus our analysis on SDS micelles to determine the transferability of results between the model compound and the actual SDS surfactant. We simulated a preassembled SDS micelle with an aggregation number of

100 in 0.8 NaCl, which are conditions where Tang et. al. first observed artifacts.[233] As illustrated in **Figure 6.1.3.2**, crystalline patches were only observed in GROMOS45A3, and the corrected GROMOS45A3KBFF produced no discernable patches. Further RDF analysis demonstrates many similarities between the solvent structure of sodium methyl sulfate and SDS micelles. Specifically, the strong peak in the Na^+-SDS(O) RDF (**Figure 6.1.3.3a**) was again observed only for the GROMOS45A3 force field, which corresponds to previous observations in the Na^+-MESU(O) RDF. Similarly, there is a minor reduction in the first solvation shell across all ionic species (**Figure 6.1.3.3c-e**), although the effect is not as pronounced as in sodium methyl sulfate. Interestingly, the Na^+-Cl^- RDF peak (**Figure 6.1.3.3b**) for both GROMOS45A3 and GROMOS53A6 is high, indicative of excessive ion pairing between Na^+ and Cl^-, but this property had no discernable effect on the SDS micelle. We suggest that this may be because the SDS micelles were simulated with a low concentration of 0.8M NaCl.

The simulation of SDS in a salt environment also provides an opportunity to examine the 3-way interactions between SDS, Na^+ and Cl^-, in terms of comparing the different implementations of 2nd-gen ion models. In the KBFF algorithm, interactions between cations and water are scaled, and only the Na^+-SPC/E scaling term was applied in our SDS simulations.[244,245] It is interesting that with only the Na^+-SPC/E scaling term, the interaction between Na^+-SDS(O) (see **Figure 3.2.3.4** for details) was indirectly modified as well, indicating that scaling the Na^+ to water interaction has some second order effect on the interactions between Na^+ and other anionic species. In contrast, the NBFIX algorithm provides additional interaction terms between specific ion pairs that scales them accordingly,[247,249] and in the context of our SDS simulations NBFIX terms for Na^+-Cl^- and Na^+-SDS(O) interactions were applied. In the context of osmotic pressure measurements, deviations from ideal behavior can be manifested when there is a differential preference for ions to interact with each other (i.e. contact ion pair) as opposed to interacting with solvent (i.e. solvent-mediated ion pair). Modulating such an effect can be achieved by either scaling attractive forces between ions which indirectly modulates the solvation shell around ions (NBFIX) or adjusting the strength of water-cation interactions, which indirectly changes the strength of other ion-ion interactions. In that sense, modifying one interaction should always have a secondary effect on the other ion pair interactions. Based on our analysis, it appears that changes in the water-cation affinity has magnified consequences compared to changes in the ion-ion affinity.

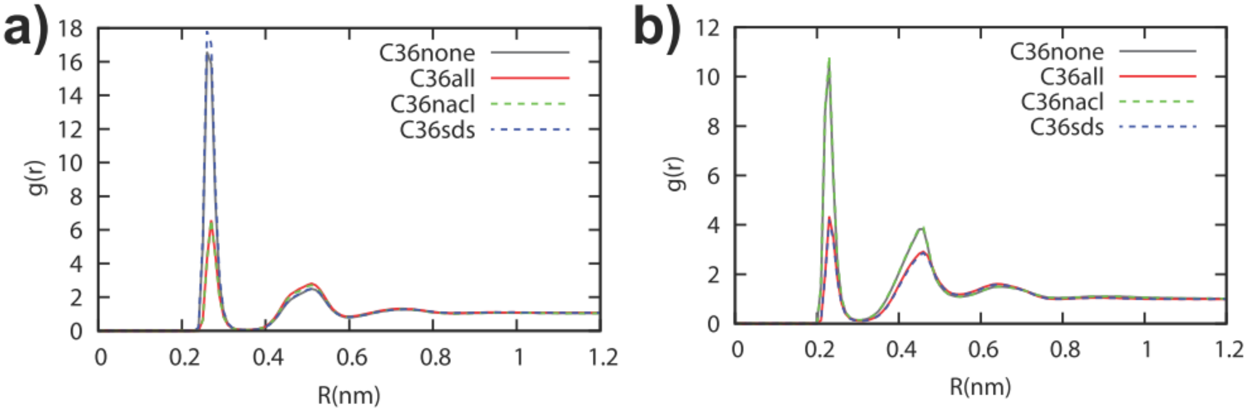


Figure 6.1.3.4: RDF of (a) Na$^+$ to Cl$^-$ and (b) Na$^+$ to SDS(O) indicates that NBFIX scaling parameters do not have second order effects on the interactions of other ion pairs.

Lastly, to determine the effect of the force field on larger SDS constructs representative of industrially-relevant concentrations, we simulated SDS micelles with aggregation number of 400 in 0.26M NaCl. Here, we limit our analysis to GROMOS45A3 which has been previously reported to produce an incorrect morphology,[233] and GROMOS45A3KBFF, which according to the Na$^+$-SDS(O) and Na$^+$-Cl$^-$ RDF (**see Figure 6.1.3.3a, b**) obtained from the 100 SDS aggregate simulation, has an equivalent behavior to the other force fields that has produced reasonable geometries.[233] As shown in **Figure 6.1.3.5**, regardless of the initial configuration, the GROMOS45A3 simulation formed a bicelle within 10 ns, with significant Na$^+$ condensation

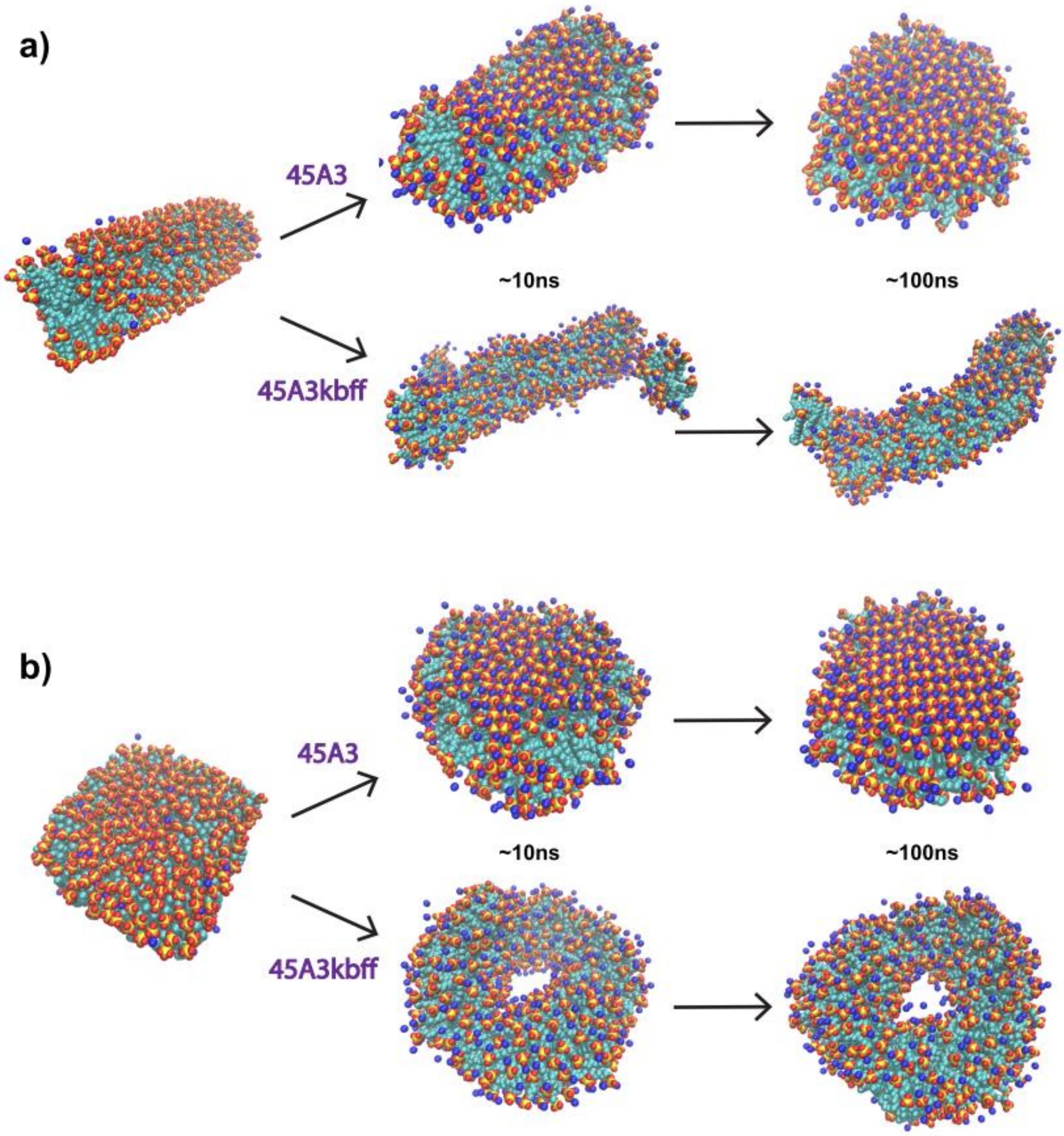


Figure 6.1.3.5: (a) Time evolution of the geometry of a SDS micelle of aggregation number 400 starting from (a) cylinder and (b) bilayer configuration. The GROMOS45A3 simulations produced incorrect bicelle structures within 10 ns, and the GROMOS45A3KBFF simulations produced correct rod-like micelles.

around the SDS head group. This structure is inconsistent with experimental data that suggests that rod-like micelles are the dominant structure in the regime between the first and second CMC.[230-232] In contrast, in the GROMOS45A3KBFF simulation, the structure formed a rod-like micelle or a toroidal micelle (a rod-like micelle looped back onto itself) within 10 ns. Thus, our results indicate that using osmotic pressure data of model compounds provide a robust method for parameter development, as it is transferrable to the simulation results of the corresponding surfactant at high aggregation number.

6.1.4 Modeling Temperature & Ionic Strength Structural Dependence

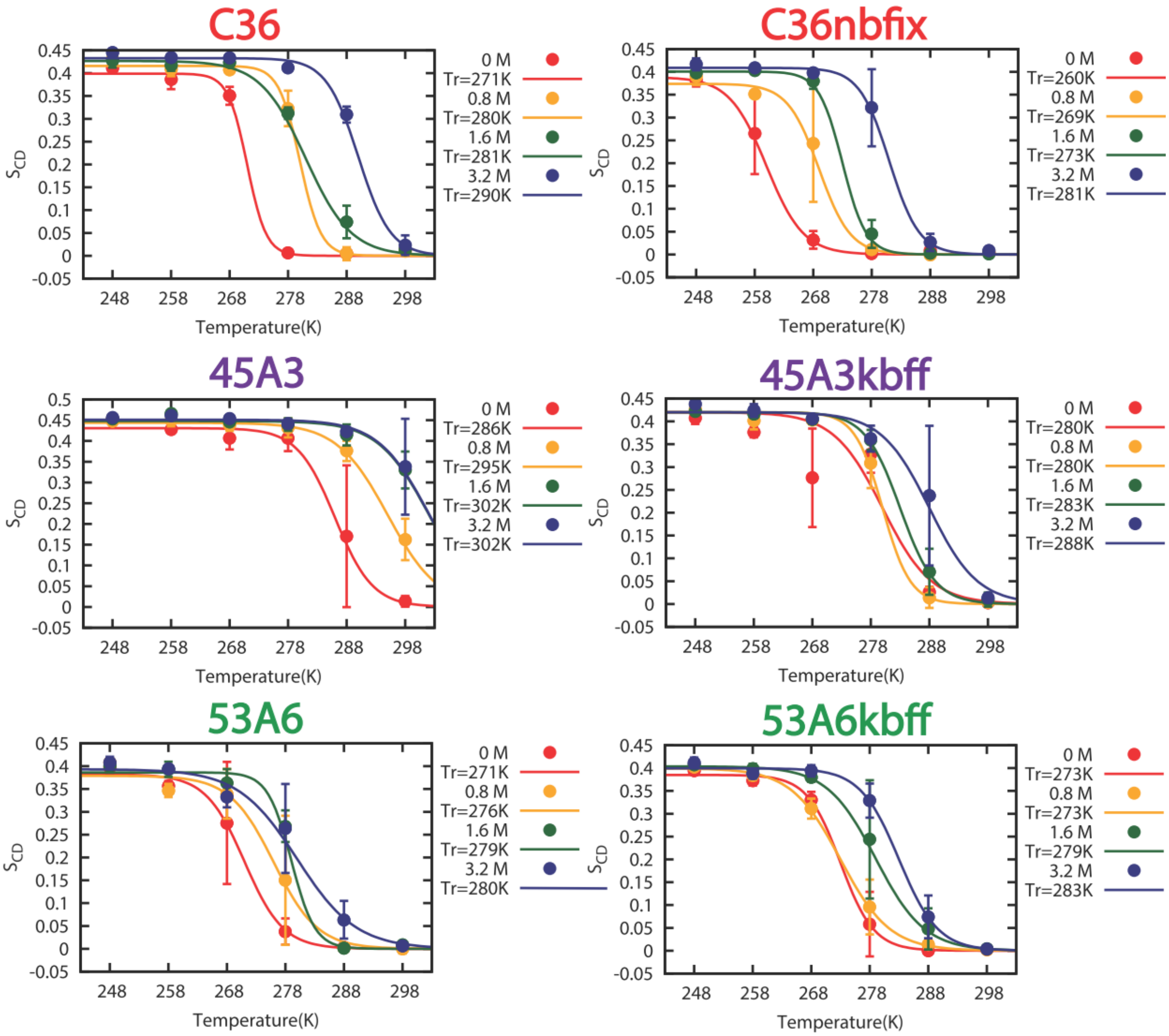


Figure 6.1.4.1: Calculated deuterium order parameter (S_{CD}) as a function of temperature and ionic strength for the 6 force fields tested. Transition temperature (T_r) from a more ordered bicelle structure ($S_{CD} > 0.35$) to a more disordered micelle structure (SCD ≈ 0) is indicated for each ionic strength.

Lastly, we analyze the effects of temperature and ionic strength on the structure of SDS micelles, and examine the effect of counterion condensation on phase transition behavior. Using a preassembled bilayer of 100 SDS molecules, we varied the temperature between 248K and 298K, and ionic strength from 0M to 3.2M. From experimental studies on anionic surfactants, it is known that increased ionic strength leads to an increase in the phase transition temperature (T_{tr}) from an ordered to disordered phase[252]. We calculated the deuterium order parameter (S_{CD}) of the hydrocarbon tails of the SDS molecules, where it is expected that a more ordered bicelle structure will have a higher S_{CD} value than a micelle structure, and the T_{tr} in our simulations can be approximated as the inflexion point for this change in S_{CD} value. As illustrated in **Figure 6.1.4.1**, the trend of ionic strength and T_{tr} is reflected across all force fields tested, although the absolute T_{tr} values and the shift in T_{tr} values between zero and high ionic strength environments do vary.

To determine the details of counterion condensation on phase transition we further analyzed the distribution of ions around the SDS head group by computing the weighted number of ions present in the first two solvation shells. Based on the RDFs calculated (data not shown), the 1st and 2nd solvation shell peak at ~0.35nm and ~0.55nm respectively, and this solvation pattern is consistent across all force fields tested. We calculated the weighted count of ions present in each shell, where ion count in the 2nd shell was scaled by a factor of 0.4 to reflect the $1/r^2$ falloff of the electrostatic contributions of the ions in the 2nd shell relative to the 1st shell, where r is the distance of the solvation shell as observed in the RDFs. Therefore, when counting the number of ions in the second shell, we scaled it by a factor of $\frac{\left(1/_{0.55}\right)^2}{\left(1/_{0.35}\right)^2} = 0.4$, in order to adjust its weight relative to the first shell. The number of ions in the first shell was not weighted, as it served as the reference point for the weighting procedure. The sum of these two counts would then give the weighted count of ions used in our analysis.

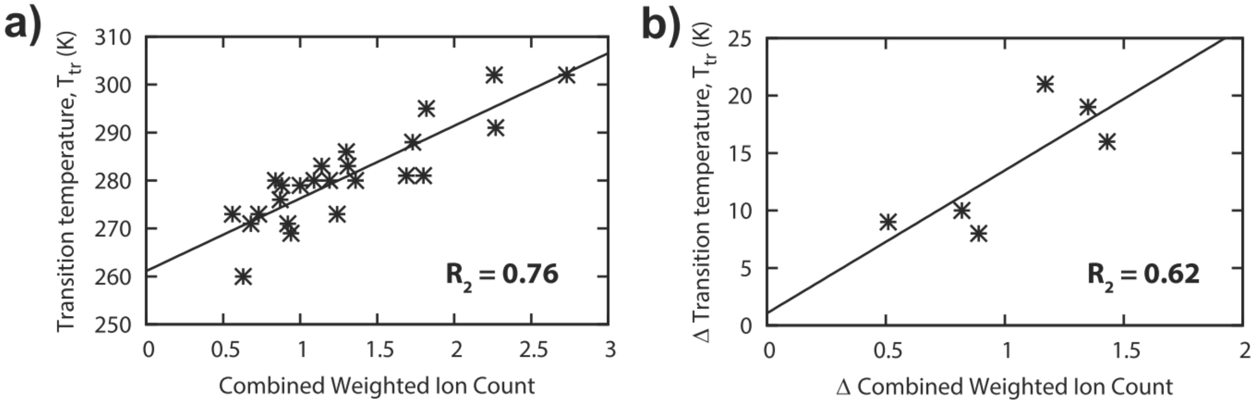


Figure 6.1.4.2: (a) The high correlation between the total ion count (N_t) and predicted transition temperature (T_{tr}), suggest that ion condensation is a primary determinant for predicting T_{tr} value. (b) Moderate correlation was observed between the difference in total ion count (ΔN_t) and the difference in phase transition temperature (ΔT_{tr}) when moving from a zero to high ionic strength environment.

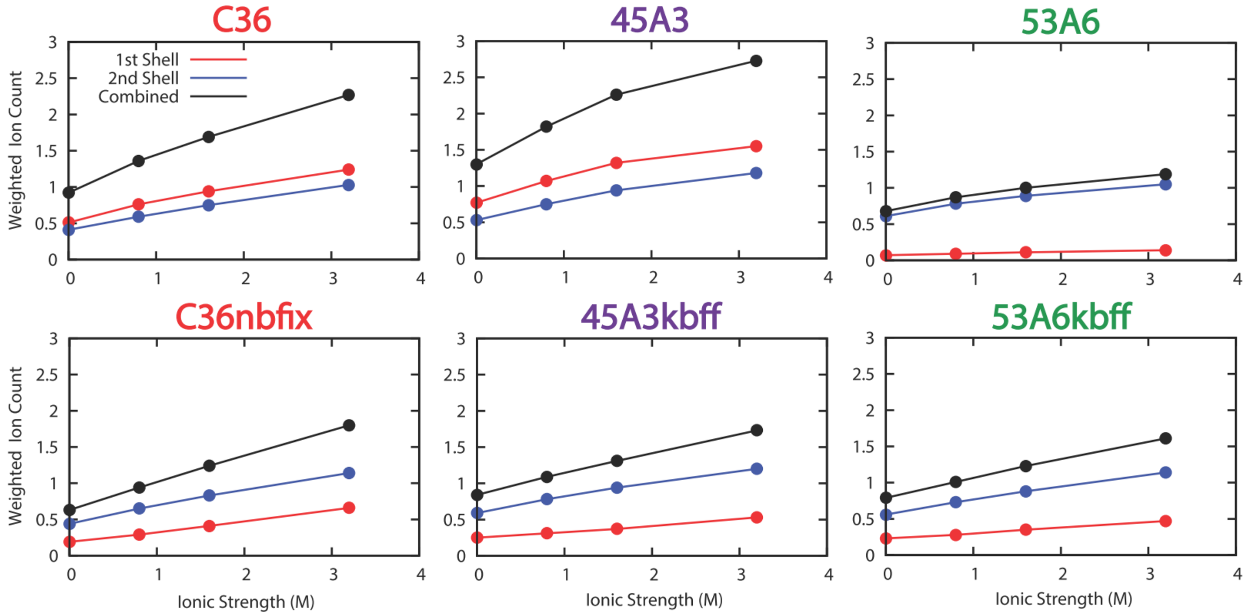


Figure 6.1.4.3: Breakdown of the weighted number of ions present in the 1st and 2nd solvation shell for all 6 force fields tested. While there are no distinct trends across 1st gen ion parameters, all 2nd gen ion parameters predict that the contribution of ions in the 2nd solvation shell outweigh those from the 1st solvation shell by a factor of 2 to 2.5.

As shown in **Figure 6.1.4.2a**, the total ion count across both shells was well correlated to the predicted T_{tr} with $R^2 = 0.77$, indicating that counterion condensation around the SDS head group is the main determinant of the shift in T_{tr}. By breaking down the contribution of counterions as a function of their solvation shells (**Figure 6.1.4.3**), there is apparently no consistency across all 1st gen ion models. Specifically, contributions from both the 1st and 2nd shell are equivalent across all ionic strength environments tested for the CHARMM36 force field, whereas the GROMOS45A3 force field predicts that 1st shell ion effects are dominant. For GROMOS53A6, there are effectively zero ions in the 1st shell, which suggests that the force field predicts that Na^+ interacts with the anionic head group of SDS exclusively via water-mediated interactions. On the other hand, all 2nd-gen ion models are qualitatively identical and predict that contributions of the 2nd shell outweigh those of the 1st shell across all ionic environments by a factor of ~2 to ~2.5. Interestingly, up to this point, both GROMOS53A6 and GROMOS53A6KBFF force fields are effectively indistinguishable based on macroscopic results, such as SDS micelle structure (formation of crystalline patches), phase transition temperature and its dependence on ionic strength. However, the two force fields are distinctly different in the ionic "microstructure" near the SDS head group. This suggests that experimental data that measures the number of ions in the 1st and 2nd solvation shell can be used as a further means of tuning and validating the finer considerations of the force field. An alternative interpretation of these results is that obtaining the correct details of the microstructure around ionic groups may not be necessary for reproducing the desired properties on the macroscopic scale.

It is also interesting to note that the width of the change in phase transition temperature (ΔT_{tr}) upon changing the ionic strength is also dependent on the force field. Specifically, CHARMM-based force fields have a ΔT_{tr} of ~20K, whereas GROMOS-based force fields have a ΔT_{tr} of ~10K. As shown in **Figure 6.1.4.2b**, we find that there is fair correlation (R^2 = 0.62) between the difference in ion count (ΔN_t) and the difference in phase transition temperature (ΔT_{tr}). While our results suggests that the differential ion condensation is contributing to the predicted ΔT_{tr}, the effects of van der Waals interactions, which should not be expected to the same across different force field families may also play a role in determining the width of the ΔT_{tr} values as a function of ionic strength.

6.1.5 Application of Parameter Validation in Other Contexts

As future applications will advance into modeling complex and concentrated ionic environments, where 3 or more ionic species are present at high concentrations, it may be constructive to contrast the different approaches used by 2nd-gen ion models. At a first glance, the parameterization strategy based on the KBFF approach may be more tractable, as only cation-water interaction terms needs to be parameterized, and so the number of additional parameters needed would scale as a factor of ~N, where N is the total number of ionic species present. While this approach has been successful in modeling SDS surfactants, it remains to be determined if a non-specific cation-water scaling term will be able to correctly modulate the subtle interactions in more complex formulations, where additional ionic species, including organic ions and multivalent ions may be present. In contrast, the NBFIX approach allows for a more precise modulation of interactions between ionic species, since there is no second order effect on related interactions. However, parameterizing NBFIX terms could scale as an order of N^2, since a pairwise potential may be needed for every permutation of ion-ion interaction. In such a situation, it may be advantageous to reduce complexity by employing a "tiered" scaling strategy, where interactions are scaled based on the interacting atom types, which is similar to the approach adopted by the GROMOS53A6 force field.[253]

Lastly, while we have shown that the importance of validating parameters against osmotic pressure and other thermodynamic measurements at finite concentration is of particular importance to SDS, and by extension to other ionic surfactants in general,[254] the basic principles behind the emphasis of local concentration of ionic species at the solute-solvent boundary may be applicable in other contexts. In simulations of biomolecular crowding[255-257] or hotspot mapping,[258-261] there can be a large quantity of cosolvents present, which may increase the effective local concentration between cosolvents and the protein. In recent advances of long timescale protein folding simulations, different force fields produce subtle differences in the folding mechanism.[262] Protein force fields have been constructed in an additive fashion using small molecule fragments (typically parameterized against dilute experimental data) for parameterizing side chains,[127] and this process may pre-bias the current generation of force fields towards dilute environments. Therefore, in protein folding simulations, where the balance between dilute (i.e. solvent exposed) and concentrated (i.e. solvent excluded) environments is key, a lack of balance between the two components may result in the current issues noted, and ongoing efforts in developing a Kirkwood-Buff inspired protein force field may resolve some issues.[263] Finally, in modeling pH effects, such as in constant pH molecular dynamics simulation, the ionic environment around titrating residues can alter the protonation state and the associated pH-dependent dynamical response. In systems with high local concentration of ionic species, such as nucleic acids,[111-113] ionic surfactants,[264,265] and charged lipids, the use of properly validated parameters for counterions may be important for reducing the occurrence of anomalous ion condensation around the titrating group.

6.1.6 Conclusion: Parameter Validation using Finite Concentration Experimental Data is Critical to Accurately Model High Concentration Environments

In this chapter, we identified that the cause of artifacts or anomalous observations reported in MD simulations of larger SDS constructs is a direct consequence of using poor parameters for modeling ionic interactions in a high concentration environment. While the global concentration of the system may be low, we discovered that the local concentration at the surfactant-solvent boundary of various SDS constructs ranges from 2 to 4 M, and this is a key consideration that needs to be realized. The most cost effective means to validate existing force fields for simulating such high concentration environments, is the use of osmotic pressure and/or other thermodynamic properties measured at finite concentration as target data for the parameterization process. By using only the properties of model compounds, specifically the ionic head groups of surfactants, we demonstrated that accurate reproduction of osmotic pressure for these model compounds translated to the correct morphology of larger SDS micelles (~400 molecules). Our investigation into the phase transition behavior of surfactants demonstrates that the total ionic strength of the simulated environment produces the expected shift in transition temperature. Furthermore, these results also suggests that macroscopic properties of SDS micelles can be insensitive to the microstructure around ionic atoms, which suggests that experimental data to distinguish between 1st and 2nd shell counterions might be useful to further validate existing parameters. Alternatively it may be interpreted that the finer details of the microstructure around ionic groups may not be necessary for accurately reproducing properties on the macroscale. Lastly, our findings on the importance of optimizing parameters for simulations in a high concentration environment may be applicable in other contexts, such as molecular crowding, hotspot mapping, protein folding, and modeling pH effects.

Chapter 7: Conclusion

7.1 Methodological Advances in Constant pH MD Simulations

In conclusion, we have developed an improved explicit solvent CPHMD framework based on the newer multi-site λ-Dynamics (MSλD) algorithm for propagating protonation states. Also known as CPHMD$^{MS\lambda D}$, it is the first viable explicit solvent CPHMD to be reported, which compared to existing implementations of explicit solvent CPHMD, the sampling in our CPHMD$^{MS\lambda D}$ framework sees a 10-fold improvement, while maintaining sufficient residency time of the physical protonation states to ensure proper solvent reorganization. In the CPHMD$^{MS\lambda D}$ framework, we performed seamless alchemical transitions between protonation and tautomeric states using MSλD, and designed a novel biasing potential to ensure that only the physical end-states are predominantly sampled. Apart from protein residues, we also developed model potentials for major nucleobases observed in both DNA and RNA, as well as additional functional groups, such as backbone phosphate and 2'OH implicated in RNA activity. In addition, we determined the proper treatment for dealing with coupled titrating systems where the identity of various residues cannot be pre-determined, which underscores the distinction between microscopic vs macroscopic (apparent) pK_a measurements.

Subsequent studies on larger full-sized proteins and nucleic acids, demonstrate the ability of CPHMD$^{MS\lambda D}$ simulations to simulate realistic pH-dependent properties of a number of model full-sized biomolecules, including HEWL, BBL and NTL9, and the lead-dependent ribozyme. Our pK_a calculations for HEWL protein are in excellent agreement with experimental values, with a RMSE of 0.84 pK_a units, and this is close to the uncertainty of 0.50 pK_a units associated with experimental measurements. Our pK_a calculations on the other model protein systems, BBL and NTL9 also provide similarly good agreement with experiments. For RNA pK_a calculations, our initial values calculated from $CPHMD^{MS\lambda D}$ simulations agree well with experimental pK_a values with an average unsigned error of 1.3 pK_a units, and the direction of the pK_a shifts for all residues in the lead-dependent ribozyme are also correctly predicted when compared to experimental data or structural considerations. Using the GAAA tetraloop and the A^{+}•C base pair of the lead-dependent ribozyme as examples, we demonstrated that $CPHMD^{MS\lambda D}$ simulations are able to model the effects that conformational dynamics and coupled titrating interactions have on the protonation equilibria of titrating residues.

Furthermore, we have also identified sampling challenges when modeling pH-dependent behavior of RNA structures. Consequently, we have enhanced the framework with pH-based replica exchange (pH-REX) sampling, which significantly improved sampling of titration and spatial coordinates, and the shuffling of conformations across pH space has the effect of decoupling interactions between titrating residues. This allows us to ameliorate some of the sampling issues related to orthogonal barriers that originate from coupled protonation equilibrium and conformational-dependent pK_a behavior, and this has the overall effect of improving accuracy from our initial results. The scalability of pH-REX sampling was also demonstrated by showing that similarly accurate pK_a values could be achieved when simulating full-sized nucleic acid systems. Finally, we highlighted that pH-REX $CPHMD^{MS\lambda D}$ simulations can be used to identify the dominant conformation of nucleic acid structures in alternate pH environments or to provide structural characterization of pH-dependent transient states, making it a useful tool to provide structural and mechanistic insight in the study of pH-dependent properties of nucleic acids.

In developing the CPHMD$^{MS\lambda D}$ framework for nucleic acids, a number of alternative protonation states had to parameterized. For the CHARMM force field, it was determined that the current CHARMM36 force field destabilizes RNA structures that have more complex structural topology beyond an A-form helix on the longer (>100 ns) timescale, and this observation is independent of the optimization of the partial charge distribution and the strength of the protonated base pairs. Our findings indicate that this was correlated to sampling the 150° to 250° region of the 2'hydroxyl dihedral phase space, which promoted the sampling of non-canonical structures. For the AMBER force field, we tested a number of solvation models, basis set and level of theories for the QM calculations used to derive the RESP charges used to describe the partial charge distribution. We discovered that there is an apparent insensitivity of interaction energies and the resulting pK_a calculations of protonated A•C$^+$ base pairs to the specifics of the QM calculations. Therefore, to maintain consistency with the standard AMBER parameterization protocol, we decided to maintain the standard gas-phase HF/6-31G* QM calculation for calculating the partial charges of alternative protonation states of nucleic acids. Lastly, our investigation into parameter validation indicates that the incorporating target data that includes finite concentration experimental data such as osmotic pressure, may be applied to further improve the accuracy of non-bonded interactions .particularly for high charge density environments, such as ionic surfactants and possibly modeling specific ion effects and nucleic acids as well.

7.2 Using CPHMD Simulations to Elucidate the Role of pH-dependent Transient States in Nucleic Acids and Proteins

Our first application of CPHMD$^{MS\lambda D}$ simulations, used in conjunction with NMR studies examined the structural characteristics of protonated Hoogsteen GC$^+$ base pairs in DNA. Our pK_a calculations indicate that the cytosine in a Hoogsteen GC$^+$ base pair is elevated to 7.1, which is in good agreement with the inferred pK_a value obtained from a N1-methyl-G variant that traps the base pair in the Hoogsteen conformation. Using both NMR data, CPHMD$^{MS\lambda D}$, we determined that transient Hoogsteen GC$^+$ base pairs are present even at physiological pH, albeit at a low population ranging from 0.1 to 0.01%, and this finding has potential implications in DNA recognition and binding by cellular factors. Moreover, we demonstrate that, at physiological pH, GC base pairs containing N1-methyl-G damage exist as a nearly equal mixture of protonated HG GC$^+$ base pairs and distorted WC-like conformers that could be specifically recognized by DNA repair enzymes in search for damaged DNA.

The role of pH-dependent transient states was later expanded to include a number of RNA systems implicated in pH-mediated RNA activity. In our analysis of protonated nucleotides across 5 different RNA the relative pK_a shifts calculated was internally consistent based on the strength of the interactions that the protonated base form with its local microenvironment, but paradoxically is not always consistent with experimental pK_a measurements. Using the U6 ISL as a precedence, we demonstrate how different pH-triggered conformational changes can alter the microscopic pK_a of each conformation, and that the apparent pK_a measured is likely to be a superposition of the pK_a values of these conformations. Based on this observation, we hypothesized that based on the elevated pK_a values of the protonated residues in the hairpin ribozyme and BWYV, there may be additional conformational states, possibly transiently populated, that are involved in its activity at physiological pH.

Using the hairpin ribozyme as an illustrative example, we used a combination of CPHMD$^{MS\lambda D}$ simulations augmented with enhanced sampling techniques to examine the details of its catalytic mechanism. Notably, the discovery of pH-mediated transient states that involve an upshifted pK$_a$ of the backbone phosphate, led to a proposed dual pathway of the hairpin ribozyme catalysis: (i) a dominant catalytic pathway involving A38/G8 as the general acid-base, which is the consensus model in the field, and (ii) a shadow catalytic pathway involving the non-bridging oxygens of the backbone phosphate. This dual pathway mechanism proposed was able to reconciled several puzzling observations, including the differential effects of mutational studies of A38 and G8 on the catalytic rate, and seemingly contradictory experimental observations, including the residual catalytic effect in an abasic A38 mutant under low pH conditions. Furthermore, we have also identified that the ground state crystallographic structure, which is best represents the catalytically active state of the hairpin ribozyme has a pK$_a$ of A38 that is does not correspond to the experimental pK$_a$ measured. This is because the hairpin ribozyme fluctuates been a dominant active state and a relaxed inactive state that is transiently populated, and this finding highlights the challenge of deconvoluting and interpreting pH-mediated mechanism particularly when dealing with conformationally flexible systems such as nucleic acids, each with their own distinct local electrostatic environment.

In the application of proteins, we utilized CPHMD$^{MS\lambda D}$ simulations to simulate the pH-dependent dynamics of a comprehensive set of SNase mutants with buried ionizable residues that have varying degrees of pK$_a$ shifts. Among our key findings is that a buried charged residue cannot be accommodated inside a purely hydrophobic pocket and that an open state structure for these “buried” residues, characterized by local solvation around the protonating site, was observed in all SNase mutants with highly shifted pK_a. At physiological pH, buried ionizable groups with large pK_a shifts have transiently populated open states, where they contribute a small but non-zero population of 1-2% at pH 7. Nevertheless, sampling these open states is a necessary condition for accurately reproducing experimental pK_a measurements, to which calculated pK_a from our explicit solvent CPHMD$^{MS\lambda D}$ simulations demonstrated good agreement, with a low average unsigned error of 1.3 pK_a units and correlation coefficient of R^2 0.78. The work we present here provides the first validation that buried ionizable residues can readily trigger pH-mediated conformational fluctuations that may be observed as transient state structures at physiological pH. Lastly, the discovery of a coupled relationship of both open and closed states and their role in recapitulating macroscopic experimental observables suggests that structural analysis of buried residues may benefit from the perspective of looking at structural pairs, as opposed to the conventional approach of a single static ground state conformation.

Using a combination of the weighted-ensemble sampling using with a novel hydration parameter and CPHMD$^{MS\lambda D}$ simulations, we further explored the role of pH-dependent transient states in elucidating the origin of an unusual non-monotonic pH-dependent absorption behavior of a recently engineered CFP mutant that features a pH-dependent shift between cyan and green fluorescent forms. In earlier experimental observations, this optical property was proposed to be controlled by the charged anionic state of its tryptophan-containing chromophore. Our calculations demonstrate that even in the presence of the stabilizing V61K mutation, the free energy cost of deprotonating the chromophore is still high and does not allow for the existence of the charged state of WasCFP even at basic pH. Instead, we propose the following explanation: The distribution between two transiently populated conformational states characterized by a partially open β-barrel with local solvation around the chromophore (that have pK_a values of 6.8 and 9.8), relative to the ground state crystallographic structure (that has pK_a of 31.7), is able to fully recapitulate both qualitatively and quantitatively the unusual non-monotonic pH-dependent properties of WasCFP. In this model, the open state is transiently populated at low pH, but reaches a population of 53% under mildly basic conditions, before losing its dominance and reverting to a transient state under highly basic conditions, and such mechanistic understanding may be used to further engineer the pH-sensitive fluorescent properties of WasCFP. Therefore, our work not only validates that tryptophan can be deprotonated in a biological system at mildly basic pH, but, more importantly, shows that pH-dependent transient conformational states are functionally relevant in proteins, and that they can tune the optical properties of fluorescent proteins. Such an outlook will have implications in the rational design of fluorescent proteins with pH-dependent optical properties.

References

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

(2) Harris, T. K.; Turner, G. J. Structural basis of perturbed pKa values of catalytic groups in enzyme active sites. *IUBMB Life* **2002**, *53*, 85-98.

(3) Dillet, V.; Dyson, H. J.; Bashford, D. Calculations of electrostatic interactions and pKas in the active site of Escherichia coli thioredoxin. *Biochemistry* **1998**, *37*, 10298-10306.

(4) Demchuk, E.; Genick, U. K.; Woo, T. T.; Getzoff, E. D.; Bashford, D. Protonation states and pH titration in the photocycle of photoactive yellow protein. *Biochemistry* **2000**, *39*, 1100-1113.

(5) Nielsen, J. E.; Mccammon, J. A. Calculating pKa values in enzyme active sites. *Protein Sci.* **2003**, *12*, 1894-1901.

(6) Wilcox, J. L.; Ahluwalia, A. K.; Bevilacqua, P. C. Charged Nucleobases and Their Potential for RNA Catalysis. *Acc. Chem. Res.* **2011**, *44*, 1270-1279.

(7) Krishnamurthy, R. Role of pK(a) of Nucleobases in the Origins of Chemical Evolution. *Acc. Chem. Res.* **2012**, *45*, 2035-2044.

(8) Shih, I. H.; Been, M. D. Involvement of a Cytosine Side Chain in Proton Transfer in the Rate-determining Step of Ribozyme Self-cleavage. *Proc. Natl. Acad. Sci. U S A* **2001**, *98*, 1489-1494.

(9) Ryder, S. P.; Oyelere, A. K.; Padilla, J. L.; Klostermeier, D.; Millar, D. P.; Strobel, S. A. Investigation of Adenosine Base Ionization in the Hairpin Ribozyme by Nucleotide Analog Interference Mapping. *RNA* **2001**, *7*, 1454-1463.

(10) Wadkins, T. S.; Shih, I.; Perrotta, A. T.; Been, M. D. A pH-sensitive RNA Tertiary Interaction Affects Self-cleavage Activity of the HDV Ribozymes in the Absence of Added Divalent Metal Ion. *J. Mol. Biol.* **2001**, *305*, 1045-1055.

(11) Ke, A.; Zhou, K.; Ding, F.; Cate, J. H.; Doudna, J. A. A Conformational Switch Controls Hepatitis Delta Virus Ribozyme Catalysis. *Nature* **2004**, *429*, 201-205.

(12) Kuzmin, Y. I.; Da Costa, C. P.; Cottrell, J. W.; Fedor, M. J. Role of an Active Site Adenine in Hairpin Ribozyme Catalysis. *J. Mol. Biol.* **2005**, *349*, 989-1010.

(13) Cerrone-Szakal, A. L.; Siegfried, N. A.; Bevilacqua, P. C. Mechanistic Characterization of the HDV Genomic Ribozyme: Solvent Isotope Effects and Proton Inventories in the Absence of Divalent Metal Ions Support C75 as the General Acid. *J. Am. Chem. Soc.* **2008**, *130*, 14504-14520.

(14) Bierzynski, A.; Kim, P. S.; Baldwin, R. L. A Salt Bridge Stabilizes the Helix Formed by Isolated C-Peptide of Rnase-A. *Proc. Natl. Acad. Sci. U S A* **1982**, *79*, 2470-2474.

(15) Shoemaker, K. R.; Kim, P. S.; Brems, D. N.; Marqusee, S.; York, E. J.; Chaiken, I. M.; Stewart, J. M.; Baldwin, R. L. Nature of the Charged-Group Effect on the Stability of the C-Peptide Helix. *Proc. Natl. Acad. Sci. U S A* **1985**, *82*, 2349-2353.

(16) Kelly, J. W. Alternative conformations of amyloidogenic proteins govern their behavior. *Curr. Opin. Struct. Biol.* **1996**, *6*, 11-17.

(17) Schaefer, M.; Van Vlijmen, H. W. T.; Karplus, M. Electrostatic contributions to molecular free energies in solution. *Adv. Protein Chem.* **1998**, *51*, 1-57.

(18) Sheinerman, F. B.; Norel, R.; Honig, B. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153-159.

(19) Warshel, A. Electrostatic Basis of Structure-Function Correlation in Proteins. *Acc. Chem. Res.* **1981**, *14*, 284-290.

(20) Hunenberger, P. H.; Helms, V.; Narayana, N.; Taylor, S. S.; McCammon, J. A. Determinants of ligand binding to cAMP-dependent protein kinase. *Biochemistry* **1999**, *38*, 2358-2366.

(21) Houck-Loomis, B.; Durney, M. A.; Salguero, C.; Shankar, N.; Nagle, J. M.; Goff, S. P.; D'Souza, V. M. An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature* **2011**, *480*, 561-U193.

(22) Webb, B. A.; Chimenti, M.; Jacobson, M. P.; Barber, D. L. Dysregulated pH: A Perfect Storm for Cancer Progression. *Nat Rev Cancer* **2011**, *11*, 671-677.

(23) Howell, E. E.; Villafranca, J. E.; Warren, M. S.; Oatley, S. J.; Kraut, J. Functional-Role of Aspartic Acid-27 in Dihydrofolate-Reductase Revealed by Mutagenesis. *Science* **1986**, *231*, 1123-1128.

(24) Rastogi, V. K.; Girvin, M. E. Structural changes linked to proton translocation by subunit c of the ATP synthase. *Nature* **1999**, *402*, 263-268.

(25) Bullough, P. A.; Hughson, F. M.; Skehel, J. J.; Wiley, D. C. Structure of Influenza Hemagglutinin at the Ph of Membrane-Fusion. *Nature* **1994**, *371*, 37-43.

(26) Nixon, P. L.; Giedroc, D. P. Energetics of a Strongly pH dependent RNA Tertiary Structure in a Frameshifting Pseudoknot. *J. Mol. Biol.* **2000**, *296*, 659-671.

(27) Bayfield, M. A.; Dahlberg, A. E.; Schulmeister, U.; Dorner, S.; Barta, A. A Conformational Change in the Ribosomal Peptidyl Transferase Center Upon Active/inactive Transition. *Proc. Natl. Acad. Sci. U S A* **2001**, *98*, 10096-10101.

(28) Muth, G. W.; Chen, L.; Kosek, A. B.; Strobel, S. A. pH-dependent Conformational Flexibility Within the Ribosomal Peptidyl Transferase Center. *RNA* **2001**, *7*, 1403-1415.

(29) Xiong, L.; Polacek, N.; Sander, P.; Bottger, E. C.; Mankin, A. pKa of Adenine 2451 in the Ribosomal Peptidyl Transferase Center Remains Elusive. *RNA* **2001**, *7*, 1365-1369.

(30) Hesslein, A. E.; Katunin, V. I.; Beringer, M.; Kosek, A. B.; Rodnina, M. V.; Strobel, S. A. Exploration of the Conserved A+C Wobble Pair Within the Ribosomal Peptidyl Transferase Center using Affinity Purified Mutant Ribosomes. *Nucleic Acids Res.* **2004**, *32*, 3760-3770.

(31) Beringer, M.; Bruell, C.; Xiong, L.; Pfister, P.; Bieling, P.; Katunin, V. I.; Mankin, A. S.; Bottger, E. C.; Rodnina, M. V. Essential Mechanisms in the Catalysis of Peptide Bond Formation on the Ribosome. *J. Biol. Chem.* **2005**, *280*, 36065-36072.

(32) Beringer, M.; Rodnina, M. V. The Ribosomal Peptidyl Transferase. *Mol. Cell* **2007**, *26*, 311-321.

(33) Abeysirigunawardena, S. C.; Chow, C. S. pH-dependent Structural Changes of Helix 69 from Escherichia Coli 23S Ribosomal RNA. *RNA* **2008**, *14*, 782-792.

(34) Sakakibara, Y.; Chow, C. S. Probing Conformational States of Modified Helix 69 in 50S Ribosomes. *J. Am. Chem. Soc.* **2011**, *133*, 8396-8399.

(35) Reiter, N. J.; Blad, H.; Abildgaard, F.; Butcher, S. E. Dynamics in the U6 RNA Intramolecular Stem-loop: A Base Flipping Conformational Change. *Biochemistry* **2004**, *43*, 13739-13747.

(36) Kim, M.; Huang, T.; Miller, J. H. Competition Between MutY and Mismatch Repair at A x C Mispairs In vivo. *J. Bacterio.* **2003**, *185*, 4626-4629.

(37) Giri, I.; Stone, M. P. Wobble dC.dA Pairing 5' to the Cationic Guanine N7 8,9-dihydro-8-(N7-guanyl)-9-hydroxyaflatoxin B1 adduct: Implications for Nontargeted AFB1 Mutagenesis. *Biochemistry* **2003**, *42*, 7023-7034.
(38) Legault, P.; Pardi, A. In situ Probing of Adenine Protonation in RNA by 13C NMR. *J. Am. Chem. Soc.* **1994**, *116*, 8390-8391.
(39) Moody, E. M.; Brown, T. S.; Bevilacqua, P. C. Simple Method for Determining Nucleobase pK(a) Values by Indirect Labeling and Demonstration of a pK(a) of Neutrality in dsDNA. *J. Am. Chem. Soc.* **2004**, *126*, 10200-10201.
(40) Liu, L.; Cottrell, J. W.; Scott, L. G.; Fedor, M. J. Direct Measurement of the Ionization State of an Essential Guanine in the Hairpin Ribozyme. *Nat. Chem. Biol.* **2009**, *5*, 351-357.
(41) Cottrell, J. W.; Scott, L. G.; Fedor, M. J. The pH Dependence of Hairpin Ribozyme Catalysis Reflects Ionization of an Active Site Adenine. *J. Biol. Chem.* **2011**, *286*, 17658-17664.
(42) Viladoms, J.; Scott, L. G.; Fedor, M. J. An Active-site Guanine Participates in glmS Ribozyme Catalysis in its Protonated State. *J. Am. Chem. Soc.* **2011**, *133*, 18388-18396.
(43) Gong, B.; Chen, J. H.; Chase, E.; Chadalavada, D. M.; Yajima, R.; Golden, B. L.; Bevilacqua, P. C.; Carey, P. R. Direct Measurement of a pKa Near Neutrality for the Catalytic Cytosine in the Genomic HDV Ribozyme using Raman Crystallography. *J. Am. Chem. Soc.* **2007**, *129*, 13335-13342.
(44) Guo, M.; Spitale, R. C.; Volpini, R.; Krucinska, J.; Cristalli, G.; Carey, P. R.; Wedekind, J. E. Direct Raman measurement of an elevated base pKa in the active site of a small ribozyme in a precatalytic conformation. *J. Am. Chem. Soc.* **2009**, *131*, 12908-12909.
(45) Korzhnev, D. M.; Religa, T. L.; Banachewicz, W.; Fersht, A. R.; Kay, L. E. A Transient and Low-Populated Protein-Folding Intermediate at Atomic Resolution. *Science* **2010**, *329*, 1312-1316.
(46) Tzeng, S. R.; Kalodimos, C. G. Allosteric inhibition through suppression of transient conformational states. *Nat. Chem. Biol.* **2013**, *9*, 462-+.
(47) Sekhar, A.; Kay, L. E. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl. Acad. Sci. U S A* **2013**, *110*, 12867-12874.
(48) Lorieau, J. L.; Louis, J. M.; Schwieters, C. D.; Bax, A. pH-triggered, activated-state conformations of the influenza hemagglutinin fusion peptide revealed by NMR. *Proc. Natl. Acad. Sci. U S A* **2012**, *109*, 19994-19999.
(49) Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. Transient Hoogsteen Base Pairs in Canonical Duplex DNA. *Nature* **2011**, *470*, 498-502.
(50) Dethoff, E. A.; Chugh, J.; Mustoe, A. M.; Al-Hashimi, H. M. Functional Complexity and Regulation Through RNA Dynamics. *Nature* **2012**, *482*, 322-330.
(51) Vallurupalli, P.; Hansen, D. F.; Stollar, E.; Meirovitch, E.; Kay, L. E. Measurement of bond vector orientations in invisible excited states of proteins. *Proc. Natl. Acad. Sci. U S A* **2007**, *104*, 18473-18477.
(52) Nikolova, E. N.; Gottardo, F. L.; Al-Hashimi, H. M. Probing Transient Hoogsteen Hydrogen Bonds in Canonical Duplex DNA Using NMR Relaxation Dispersion and Single-Atom Substitution. *J. Am. Chem. Soc.* **2012**, *134*, 3667-3670.
(53) Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc. Natl. Acad. Sci. U S A* **2011**, *108*, 16247-16252.

(54) Krasovska, M. V.; Sefcikova, J.; Spackova, N.; Sponer, J.; Walter, N. G. Structural dynamics of precursor and product of the RNA enzyme from the hepatitis delta virus as revealed by molecular dynamics simulations. *J. Mol. Biol.* **2005**, *351*, 731-748.
(55) Krasovska, M. V.; Sefcikova, J.; Reblova, K.; Schneider, B.; Walter, N. G.; Sponer, J. Cations and hydration in catalytic RNA: Molecular dynamics of the hepatitis delta virus ribozyme. *Biophys. J.* **2006**, *91*, 626-638.
(56) Ditzler, M. A.; Sponer, J.; Walter, N. G. Molecular Dynamics Suggest Multifunctionality of an Adenine Imino Group in Acid-base Catalysis of the Hairpin Ribozyme. *RNA* **2009**, *15*, 560-575.
(57) Mlýnský, V.; Banás, P.; Hollas, D.; Réblová, K.; Walter, N. G.; Sponer, J.; Otyepka, M. Extensive Molecular Dynamics Simulations Showing that Canonical G8 and Protonated A38H+ Forms are Most Consistent with Crystal Structures of Hairpin Ribozyme. *J. Phys. Chem. B.* **2010**, *114*, 6642-6652.
(58) Veeraraghavan, N.; Bevilacqua, P. C.; Hammes-Schiffer, S. Long-Distance Communication in the HDV Ribozyme: Insights from Molecular Dynamics and Experiments. *J. Mol. Biol.* **2010**, *402*, 278-291.
(59) Bashford, D. Macroscopic electrostatic models for protonation states in proteins. *Front. Biosci.* **2004**, *9*, 1082-1099.
(60) Tang, C. L.; Alexov, E.; Pyle, A. M.; Honig, B. Calculation of pKas in RNA: On the Structural Origins and Functional Roles of Protonated Nucleotides. *J. Mol. Biol.* **2007**, *366*, 1475-1496.
(61) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. Prediction of pH-dependent properties of proteins. *J. Mol. Biol.* **1994**, *238*, 415-436.
(62) You, T. J.; Bashford, D. Conformation and hydrogen ion titration of proteins: a continuum electrostatic model with conformational flexibility. *Biophys. J.* **1995**, *69*, 1721-1733.
(63) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. *Biophys. J.* **2002**, *83*, 1731-1748.
(64) Russell, S. T.; Warshel, A. Calculations of electrostatic energies in proteins. The energetics of ionized groups in bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* **1985**, *185*, 389-404.
(65) Lee, F. S.; Chu, Z. T.; Warshel, A. Microscopic and Semimicroscopic Calculations of Electrostatic Energies in Proteins by the Polaris and Enzymix Programs. *J. Comput. Chem.* **1993**, *14*, 161-185.
(66) Warshel, A.; Sussman, F.; King, G. Free-Energy of Charges in Solvated Proteins - Microscopic Calculations Using a Reversible Charging Process. *Biochemistry* **1986**, *25*, 8368-8372.
(67) Sham, Y. Y.; Chu, Z. T.; Warshel, A. Consistent calculations of pKa's of ionizable residues in proteins: semi-microscopic and microscopic approaches. *J. Phys. Chem. B.* **1997**, *101*, 4458-4472.
(68) Mongan, J.; Case, D. A. Biomolecular simulations at constant pH. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157-163.
(69) Burgi, R.; Kollman, P. A.; van Gunsteren, W. F. Simulating proteins at constant pH: an approach combining molecular dynamics and Monte Carlo simulation. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 469-480.
(70) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH Molecular Dynamics Using Stochastic Titration. *J. Chem. Phys.* **2002**, *117*, 4184-4200.

(71) Machuqueiro, M.; Baptista, A. M. Constant-pH molecular dynamics with ionic strength effects: protonation-conformation coupling in decalysine. *J. Phys. Chem. B.* **2006**, *110*, 2927-2933.
(72) Baptista, A. M.; Machuqueiro, M. Acidic range titration of HEWL using a constant-pH molecular dynamics method. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 289-298.
(73) Im, W. P.; Lee, M. S.; Brooks, C. L., III. Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691-1702.
(74) Chen, J. H.; Im, W. P.; Brooks, C. L., III. Balancing solvation and intramolecular interactions: Toward a consistent generalized born force field. *J. Am. Chem. Soc.* **2006**, *128*, 3728-3736.
(75) Dlugosz, M.; Antosiewicz, J. M. Constant-pH molecular dynamics simulations: a test case of succinic acid. *Chem. Phys.* **2004**, *302*, 161-170.
(76) Dlugosz, M.; Antosiewicz, J. M.; Robertson, A. D. Constant-pH molecular dynamics study of protonation-structure relationship in a heptapeptide derived from ovomucoid third domain. *Phys. Rev. E* **2004**, *69*, 021915.
(77) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* **2004**, *25*, 2038-2048.
(78) Williams, S. L.; de Oliveira, C. A.; McCammon, J. A. Coupling Constant pH Molecular Dynamics with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2010**, *6*, 560-568.
(79) Meng, Y. L.; Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Biomolecules Using a Discrete Protonation Model. *J. Chem. Theory Comput.* **2010**, *6*, 1401-1412.
(80) Swails, J. M.; Roitberg, A. E. Enhancing Conformation and Protonation State Sampling of Hen Egg White Lysozyme Using pH Replica Exchange Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 4393-4404.
(81) Sabri Dashti, D.; Meng, Y.; Roitberg, A. E. pH-replica exchange molecular dynamics in proteins using a discrete protonation method. *J. Phys. Chem. B.* **2012**, *116*, 8805-8811.
(82) Olsson, M. H. M.; Warshel, A. Monte Carlo simulations of proton pumps: On the working principles of the biological valve that controls proton pumping in cytochrome c oxidase. *Proc. Natl. Acad. Sci. U S A* **2006**, *103*, 6500-6505.
(83) Aaqvist, J.; Warshel, A. Simulation of enzyme reactions using valence bond force fields and other hybrid quantum/classical approaches. *Chem. Rev.* **1993**, *93*, 2523-2544.
(84) Baptista, A. M.; Martel, P. J.; Petersen, S. B. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins: Struct., Funct., Bioinf.* **1997**, *27*, 523.
(85) Borjesson, U.; Hunenberger, P. H. Explicit-solvent molecular dynamics simulation at constant pH: Methodology and application to small amines. *J. Chem. Phys.* **2001**, *114*, 9706-9719.
(86) Kong, X.; Brooks, C. L., III. Lambda-Dynamics-A New Approach to Free-Energy Calculations. *J. Chem. Phys.* **1996**, *105*, 2414-2423.
(87) Knight, J. L.; Brooks, C. L., III. Lambda-dynamics free energy simulation methods. *J. Comput. Chem.* **2009**, *30*, 1692-1700.
(88) Guo, Z.; Brooks, C. L., III; Kong, X. Efficient and flexible algorithm for free energy calculations using the λ-dynamics approach. *J. Phys. Chem. B.* **1998**, *102*, 2032-2036.
(89) Lee, M. S.; Salsbury, F. R.; Brooks, C. L., III. Constant-pH Molecular Dynamics Using Continuous Titration Coordinates. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738-752.

(90) Khandogin, J.; Brooks, C. L., III. Constant pH Molecular Dynamics with Proton Tautomerism. *Biophys. J.* **2005**, *89*, 141-157.

(91) Khandogin, J.; Brooks, C. L., III. Toward the accurate first-principles prediction of ionization equilibria in proteins. *Biochemistry* **2006**, *45*, 9363-9373.

(92) Khandogin, J.; Chen, J.; Brooks, C. L., III. Exploring Atomistic Details of pH-dependent Peptide Folding. *Proc. Natl. Acad. Sci. U S A* **2006**, *103*, 18546-18550.

(93) Khandogin, J.; Raleigh, D. P.; Brooks, C. L., III. Folding Intermediate in the Villin Headpiece Domain Arises From Disruption of a N-terminal Hydrogen-bonded Network. *J. Am. Chem. Soc.* **2007**, *129*, 3056-3057.

(94) Khandogin, J.; Brooks, C. L., III. Linking Folding with Aggregation in Alzheimer's Beta-amyloid Peptides. *Proc. Natl. Acad. Sci. U S A* **2007**, *104*, 16880-16885.

(95) Zhang, B. W.; Brunetti, L.; Brooks, C. L., III. Probing pH-dependent Dissociation of HdeA Dimers. *J. Am. Chem. Soc.* **2011**, *133*, 19393-19398.

(96) Shen, J. K. Uncovering specific electrostatic interactions in the denatured states of proteins. *Biophys. J.* **2010**, *99*, 924-932.

(97) Wallace, J. A.; Shen, J. K. Unraveling A Trap-and-Trigger Mechanism in the pH-Sensitive Self-Assembly of Spider Silk Proteins. *J. Phys. Chem. Lett.* **2012**, *3*, 658–662.

(98) Law, S. M.; Zhang, B. W.; Brooks, C. L., III. pH-sensitive residues in the p19 RNA silencing suppressor protein from carnation Italian ringspot virus affect siRNA binding stability. *Protein Sci.* **2013**, *22*, 595-604.

(99) Dlugosz, M.; Antosiewicz, J. M. Effects of solute-solvent proton exchange on polypeptide chain dynamics: a constant-pH molecular dynamics study. *J. Phys. Chem. B.* **2005**, *109*, 13777-13784.

(100) Machuqueiro, M.; Baptista, A. M. The pH-dependent Conformational States of Kyotorphin: A Constant-pH Molecular Dynamics Study. *Biophys. J.* **2007**, *92*, 1836-1845.

(101) Campos, S. R.; Machuqueiro, M.; Baptista, A. M. Constant-pH molecular dynamics simulations reveal a beta-rich form of the human prion protein. *J. Phys. Chem. B.* **2010**, *114*, 12692-12700.

(102) Arthur, E. J.; Yesselman, J. D.; Brooks, C. L., III. Predicting extreme pK(a) shifts in staphylococcal nuclease mutants with constant pH molecular dynamics. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3276-3286.

(103) Wallace, J. A.; Shen, J. K. Continuous Constant pH Molecular Dynamics in Explicit Solvent with pH-Based Replica Exchange. *J. Chem. Theory Comput.* **2011**, *7*, 2617-2629.

(104) Wang, W. Z.; Chu, X. P.; Li, M. H.; Seeds, J.; Simon, R. P.; Xiong, Z. G. Modulation of acid-sensing ion channel currents, acid-induced increase of intracellular Ca2+, and acidosis-mediated neuronal injury by intracellular pH. *J. Biol. Chem.* **2006**, *281*, 29369-29378.

(105) Hesselager, M.; Timmermann, D. B.; Ahring, P. K. pH dependency and desensitization kinetics of heterologously expressed combinations of acid-sensing ion channel subunits. *J. Biol. Chem.* **2004**, *279*, 11006-11015.

(106) Berdiev, B. K.; Mapstone, T. B.; Markert, J. M.; Gillespie, G. Y.; Lockhart, J.; Fuller, C. M.; Benos, D. J. pH alterations "reset" Ca2+ sensitivity of brain Na+ channel 2, a degenerin/epithelial Na+ ion channel, in planar lipid bilayers. *J. Biol. Chem.* **2001**, *276*, 38755-38761.

(107) Damaghi, M.; Bippes, C.; Koster, S.; Yildiz, O.; Mari, S. A.; Kuhlbrandt, W.; Muller, D. J. pH-Dependent Interactions Guide the Folding and Gate the Transmembrane Pore of the beta-Barrel Membrane Protein OmpG. *J. Mol. Biol.* **2010**, *397*, 878-882.

(108) Gaillard, T.; Case, D. A. Evaluation of DNA Force Fields in Implicit Solvation. *J. Chem. Theory Comput.* **2011**, *7*, 3181-3198.
(109) Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmuller, H. Constant pH Molecular Dynamics in Explicit Solvent with lambda-Dynamics. *J. Chem. Theory Comput.* **2011**, *7*, 1962-1978.
(110) Swails, J. M.; York, D. M.; Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J. Chem. Theory Comput.* **2014**, *10*, 1341-1352.
(111) Goh, G. B.; Knight, J. L.; Brooks, C. L., III. Constant pH molecular dynamics simulations of nucleic acids in explicit solvent. *J. Chem. Theory Comput.* **2012**, *8*, 36-46.
(112) Goh, G. B.; Knight, J. L.; Brooks, C. L., III. pH-dependent dynamics of complex RNA macromolecules. *J. Chem. Theory Comput.* **2013**, *9*, 935-943.
(113) Goh, G. B.; Knight, J. L.; Brooks, C. L., III. Toward accurate prediction of the protonation equilibrium of nucleic acids. *J. Phys. Chem. Lett.* **2013**, *4*, 760-766.
(114) Goh, G. B.; Hulbert, B. S.; Zhou, H.; Brooks, C. L., III. Constant pH Molecular Dynamics of Proteins in Explicit Solvent with Proton Tautomerism. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 1319-1331.
(115) Sripathi, K.; Goh, G. B.; Dickson, A.; Walter, N. G.; Brooks, C. L., III. Expanding the Repertoire of Constant pH MD Simulations of Nucleic Acids. *manuscript in preparation*.
(116) Goh, G. B.; Sripathi, K.; Dickson, A.; Walter, N. G.; Brooks, C. L., III. What You See is Not Exactly What You Get: pH-Dependent Regulation of Hairpin Ribozyme Catalysis. *manuscript in preparation*.
(117) Nikolova, E. N.; Goh, G. B.; Brooks, C. L., III; Al-Hashimi, H. M. Characterizing the Protonation State of Cytosine in Transient G.C Hoogsteen Base Pairs in Duplex DNA. *J. Am. Chem. Soc.* **2013**, *135*, 6766-6769.
(118) Dickson, A.; Brooks, C. L., III. WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm. *J. Phys. Chem. B.* **2014**, *118*, 3532-3542.
(119) Goh, G. B.; Laricheva, E. N.; Brooks, C. L., III. Uncovering pH-dependent transient states of proteins with buried ionizable residues. *J. Am. Chem. Soc.* **2014**, *136*, 8496-8499.
(120) Laricheva, E. N.; Goh, G. B.; Dickson, A.; Brooks, C. L., III. pH-dependent transient conformational states control optical properties in cyan fluorescent protein. *J. Am. Chem. Soc.* **2015**, *137*, 2892-2900.
(121) Knight, J. L.; Brooks, C. L., III. Applying Efficient Implicit Non-geometric Constraints in Alchemical Free Energy Simulations. *J. Comput. Chem.* **2011**, *32*, 3423-3432.
(122) Knight, J. L.; Brooks, C. L., III. Multisite λ Dynamics for Simulated Structure–Activity Relationship Studies. *J. Chem. Theory Comput.* **2011**, *7*, 2728-2739.
(123) Shi, C. Y.; Wallace, J. A.; Shen, J. K. Thermodynamic Coupling of Protonation and Conformational Equilibria in Proteins: Theory and Simulation. *Biophys. J.* **2012**, *102*, 1590-1597.
(124) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545-1614.

(125) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926-935.

(126) Feig, M.; Karanicolas, J.; Brooks, C. L., III. MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Mol. Graphics Modell.* **2004**, *22*, 377-395.

(127) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* **1998**, *102*, 3586-3616.

(128) Hoogstraten, C. G.; Legault, P.; Pardi, A. NMR Solution Structure of the Lead-dependent Ribozyme: Evidence for Dynamics in RNA Catalysis. *J. Mol. Biol.* **1998**, *284*, 337-350.

(129) Denning, E. J.; Priyakumar, U. D.; Nilsson, L.; Mackerell, A. D., Jr. Impact of 2'-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *J. Comput. Chem.* **2011**, *32*, 1929-1943.

(130) Best, R. B.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Inclusion of many-body effects in the additive CHARMM protein CMAP potential results in enhanced cooperativity of α-helix and β-hairpin formation. *Biophys. J.* **2012**, *103*, 1045-1051.

(131) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the additive CHARMM all-atom protein force field rargeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257-3273.

(132) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327-341.

(133) Cheatham, T. E.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. Molecular-Dynamics Simulations on Solvated Biomolecular Systems - the Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, Rna, and Proteins. *J. Am. Chem. Soc.* **1995**, *117*, 4193-4194.

(134) Schreiber, H.; Steinhauser, O. Cutoff Size Does Strongly Influence Molecular-Dynamics Results on Solvated Polypeptides. *Biochemistry* **1992**, *31*, 5856-5860.

(135) Steinbach, P. J.; Brooks, B. R. New Spherical-Cutoff Methods for Long-Range Forces in Macromolecular Simulation. *J. Comput. Chem.* **1994**, *15*, 667-683.

(136) Beck, D. A. C.; Armen, R. S.; Daggett, V. Cutoff Size Need Not Strongly Influence Molecular Dynamics Results for Solvated Polypeptides. *Biochemistry* **2005**, *44*, 609-616.

(137) Norberg, J.; Nilsson, L. On the truncation of long-range electrostatic interactions in DNA. *Biophys. J.* **2000**, *79*, 1537-1553.

(138) Onufriev, A.; Case, D. A.; Ullmann, G. M. A novel view of pH titration in biomolecules. *Biochemistry* **2001**, *40*, 3413-3419.

(139) Klingen, A. R.; Bombarda, E.; Ullmann, G. M. Theoretical investigation of the behavior of titratable groups in proteins. *Photochem. Photobiol. Sci.* **2006**, *5*, 588-596.

(140) Bashford, D.; Karplus, M. Multiple-Site Titration Curves of Proteins - an Analysis of Exact and Approximate Methods for Their Calculation. *J. Phys. Chem.* **1991**, *95*, 9556-9561.

(141) Halgren, T. A. Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490-519.

(142) Izatt, R. M.; Christensen, J. J.; Rytting, J. H. Sites and thermodynamic quantities associated with proton and metal ion interaction with ribonucleic acid, deoxyribonucleic acid, and their constituent bases, nucleosides, and and nucleotides. *Chem. Rev.* **1971**, *71*, 439-481.

(143) Khavrutskii, I. V.; Price, D. J.; Lee, J.; Brooks, C. L., III. Conformational change of the methionine 20 loop of Escherichia coli dihydrofolate reductase modulates pKa of the bound dihydrofolate. *Protein Sci.* **2007**, *16*, 1087-1100.

(144) Alberty, R. A.; Smith, R. M.; Bock, R. M. The apparent ionization constants of the adenosinephosphates and related compounds. *J. Biol. Chem.* **1951**, *193*, 425-434.

(145) Cavalieri, L. F. Studies on the Structure of Nucleic Acids. VII. On the Identification of the Isomeric Cytidylic and Adenylic Acids1. *J. Am. Chem. Soc.* **1953**, *75*, 5268-5270.

(146) Nielsen, J. E.; Gunner, M. R.; Garcia-Moreno, E. B. The pK(a) Cooperative: A Collaborative Effort to Advance Structure-based Calculations of pK(a) Values and Electrostatic Effects in Proteins. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3249-3259.

(147) Legault, P.; Pardi, A. Unusual Dynamics and pKa Shift at the Active Site of a Lead-dependent Ribozyme. *J. Am. Chem. Soc.* **1997**, *119*, 6621-6628.

(148) Ke, A.; Ding, F.; Batchelor, J. D.; Doudna, J. A. Structural roles of monovalent cations in the HDV ribozyme. *Structure* **2007**, *15*, 281-287.

(149) Banas, P.; Walter, N. G.; Sponer, J.; Otyepka, M. Protonation States of the Key Active Site Residues and Structural Dynamics of the glmS Riboswitch As Revealed by Molecular Dynamics. *J. Phys. Chem. B.* **2010**, *114*, 8701-8712.

(150) Legault, P.; Hoogstraten, C. G.; Metlitzky, E.; Pardi, A. Order, dynamics and metal-binding in the lead-dependent ribozyme. *J. Mol. Biol.* **1998**, *284*, 325-335.

(151) Zheng, L.; Chen, M.; Yang, W. Random Walk in Orthogonal Space to Achieve Efficient Free-energy Simulation of Complex Systems. *Proc. Natl. Acad. Sci. U S A* **2008**, *105*, 20227-20232.

(152) Woese, C. R.; Winker, S.; Gutell, R. R. Architecture of Ribosomal-Rna - Constraints on the Sequence of Tetra-Loops. *Proc. Natl. Acad. Sci. U S A* **1990**, *87*, 8467-8471.

(153) Hoogstraten, C. G.; Wank, J. R.; Pardi, A. Active site dynamics in the lead-dependent ribozyme. *Biochemistry* **2000**, *39*, 9951-9958.

(154) Jucker, F. M.; Heus, H. A.; Yip, P. F.; Moors, E. H. M.; Pardi, A. A Network of Heterogeneous Hydrogen Bonds in GNRA Tetraloops. *J. Mol. Biol.* **1996**, *264*, 968-980.

(155) Menger, M.; Eckstein, F.; Porschke, D. Dynamics of the RNA Hairpin GNRA Tetraloop. *Biochemistry* **2000**, *39*, 4500-4507.

(156) Zhang, Y. F.; Zhao, X.; Mu, Y. G. Conformational Transition Map of an RNA GCAA Tetraloop Explored by Replica-Exchange Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2009**, *5*, 1146-1154.

(157) Deng, Y. Q.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B.* **2009**, *113*, 2234-2246.

(158) Brandsdal, B. O.; Osterberg, F.; Almlof, M.; Feierberg, I.; Luzhkov, V. B.; Aqvist, J. Free energy calculations and ligand binding. *Adv. Protein Chem.* **2003**, *66*, 123-158.

(159) Zheng, L. Q.; Yang, W. Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering. *J. Chem. Theory Comput.* **2012**, *8*, 810-823.

(160) Nozaki, Y.; Tanford, C. Examination of Titration Behavior. *Methods Enzymol.* **1967**, *11*, 715-734.

(161) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. Electrostatic Calculations of Side-Chain Pk(a) Values in Myoglobin and Comparison with Nmr Data for Histidines. *Biochemistry* **1993**, *32*, 8045-8056.
(162) Tanford, C.; Roxby, R. Interpretation of protein titration curves. Application to lysozyme. *Biochemistry* **1972**, *11*, 2192-2198.
(163) Parsons, S. M.; Raftery, M. A. Ionization behavior of the catalytic carboxyls of lysozyme. Effects of ionic strength. *Biochemistry* **1972**, *11*, 1623-1629.
(164) Kuramitsu, S.; Ikeda, K.; Hamaguchi, K.; Fujio, H.; Amano, T. Ionization constants of Glu 35 and Asp 52 in hen, turkey, and human lysozymes. *J. Biochem.* **1974**, *76*, 671-683.
(165) Kuramitsu, S.; Ikeda, K.; Hamaguchi, K. Participation of the catalytic carboxyls, Asp 52 and Glu 35, and Asp 101 in the binding of substrate analogues to hen lysozyme. *J. Biochem.* **1975**, *77*, 291-301.
(166) Takahashi, T.; Nakamura, H.; Wada, A. Electrostatic forces in two lysozymes: calculations and measurements of histidine pKa values. *Biopolymers* **1992**, *32*, 897-909.
(167) Bartik, K.; Redfield, C.; Dobson, C. M. Measurement of the Individual Pk(a) Values of Acidic Residues of Hen and Turkey Lysozymes by 2-Dimensional H-1-Nmr. *Biophys. J.* **1994**, *66*, 1180-1184.
(168) 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. Remeasuring HEWL pK(a) values by NMR spectroscopy: methods, analysis, accuracy, and implications for theoretical pK(a) calculations. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 685-702.
(169) Machuqueiro, M.; Baptista, A. M. Is the prediction of pKa values by constant-pH molecular dynamics being hindered by inherited problems? *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3437-3447.
(170) Wallace, J. A.; Shen, J. K. Charge-leveling and proper treatment of long-range electrostatics in all-atom molecular dynamics at constant pH. *J. Chem. Phys.* **2012**, *137*, 184105.
(171) Arbely, E.; Rutherford, T. J.; Sharpe, T. D.; Ferguson, N.; Fersht, A. R. Downhill versus Barrier-Limited Folding of BBL 1: Energetic and Structural Perturbation Effects upon Protonation of a Histidine of Unusually Low pK(a). *J. Mol. Biol.* **2009**, *387*, 986-992.
(172) Kuhlman, B.; Luisi, D. L.; Young, P.; Raleigh, D. P. pK(a) values and the pH dependent stability of the N-terminal domain of L9 as probes of electrostatic interactions in the denatured state. Differentiation between local and nonlocal interactions. *Biochemistry* **1999**, *38*, 4896-4903.
(173) Isom, D. G.; Castaneda, C. A.; Cannon, B. R.; Garcia-Moreno, B. Large shifts in pKa values of lysine residues buried inside a protein. *Proc. Natl. Acad. Sci. U S A* **2011**, *108*, 5260-5265.
(174) Isom, D. G.; Cannon, B. R.; Castaneda, C. A.; Robinson, A.; Garcia-Moreno, B. High tolerance for ionizable residues in the hydrophobic interior of proteins. *Proc. Natl. Acad. Sci. U S A* **2008**, *105*, 17784-17788.
(175) Fafarman, A. T.; Sigala, P. A.; Schwans, J. P.; Fenn, T. D.; Herschlag, D.; Boxer, S. G. Quantitative, directional measurement of electric field heterogeneity in the active site of ketosteroid isomerase. *Proc. Natl. Acad. Sci. U S A* **2012**, *109*, E299-E308.
(176) Nixon, P. L.; Cornish, P. V.; Suram, S. V.; Giedroc, D. P. Thermodynamic analysis of conserved loop-stem interactions in P1-P2 frameshifting RNA pseudoknots from plant Luteoviridae. *Biochemistry* **2002**, *41*, 10665-10674.

(177) Nixon, P. L.; Rangan, A.; Kim, Y. G.; Rich, A.; Hoffman, D. W.; Hennig, M.; Giedroc, D. P. Solution structure of a luteoviral P1-P2 frameshifting mRNA pseudoknot. *J. Mol. Biol.* **2002**, *322*, 621-633.
(178) Cornish, P. V.; Hennig, M.; Giedroc, D. P. A loop 2 cytidine-stem 1 minor groove interaction as a positive determinant for pseudoknot-stimulated -1 ribosomal frameshifting. *Proc. Natl. Acad. Sci. U S A* **2005**, *102*, 12694-12699.
(179) Denning, E. J.; MacKerell, A. D. Intrinsic Contribution of the 2 '-Hydroxyl to RNA Conformational Heterogeneity. *J. Am. Chem. Soc.* **2012**, *134*, 2800-2806.
(180) Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E., III; Jurecka, P. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886-2902.
(181) Banas, P.; Hollas, D.; Zgarbova, M.; Jurecka, P.; Orozco, M.; Cheatham, T. E., III; Sponer, J.; Otyepka, M. Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins. *J. Chem. Theory Comput.* **2010**, *6*, 3836-3849.
(182) Krepl, M.; Zgarbova, M.; Stadlbauer, P.; Otyepka, M.; Banas, P.; Koca, J.; Cheatham, T. E.; Jurecka, P.; Sponer, J. Reference Simulations of Noncanonical Nucleic Acids with Different chi Variants of the AMBER Force Field: Quadruplex DNA, Quadruplex RNA, and Z-DNA. *J. Chem. Theory Comput.* **2012**, *8*, 2506-2520.
(183) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the Multimolecule and Multiconformational Resp Methodology to Biopolymers - Charge Derivation for DNA, Rna, and Proteins. *J. Comput. Chem.* **1995**, *16*, 1357-1377.
(184) Klein, D. J.; Been, M. D.; Ferre-D'Amare, A. R. Essential role of an active-site guanine in glmS ribozyme catalysis. *J. Am. Chem. Soc.* **2007**, *129*, 14858-+.
(185) Ritchie, D. B.; Foster, D. A. N.; Woodside, M. T. Programmed-1 frameshifting efficiency correlates with RNA pseudoknot conformational plasticity, not resistance to mechanical unfolding. *Proc. Natl. Acad. Sci. U S A* **2012**, *109*, 16167-16172.
(186) Korzhnev, D. M.; Orekhov, V. Y.; Kay, L. E. Off-resonance R1(p) NMR studies of exchange dynamics in proteins with low spin-lock fields: An application to a fyn SH3 domain. *J. Am. Chem. Soc.* **2005**, *127*, 713-721.
(187) Palmer, A. G.; Massi, F. Characterization of the dynamics of biomacromolecules using rotating-frame spin relaxation NMR spectroscopy. *Chem. Rev.* **2006**, *106*, 1700-1719.
(188) Wang, A. H. J.; Ughetto, G.; Quigley, G. J.; Hakoshima, T.; Vandermarel, G. A.; Vanboom, J. H.; Rich, A. The Molecular-Structure of a DNA Triostin-a Complex. *Science* **1984**, *225*, 1115-1121.
(189) Asensio, J. L.; Lane, A. N.; Dhesi, J.; Bergqvist, S.; Brown, T. The contribution of cytosine protonation to the stability of parallel DNA triple helices. *J. Mol. Biol.* **1998**, *275*, 811-822.
(190) Nair, D. T.; Johnson, R. E.; Prakash, S.; Prakash, L.; Aggarwal, A. K. Replication by human DNA polymerase-iota occurs by Hoogsteen base-pairing. *Nature* **2004**, *430*, 377-380.
(191) Wang, J. M. Hoogsteen base-pairing in DNA replication? *Nature* **2005**, *437*, E6-E7.
(192) Nair, D. T.; Johnson, R. E.; Prakash, L.; Prakash, S.; Aggarwal, A. K. Human DNA polymerase iota incorporates dCTP opposite template G via a G.C plus hoogsteen base pair. *Structure* **2005**, *13*, 1569-1577.

(193) Patikoglou, G. A.; Kim, J. L.; Sun, L. P.; Yang, S. H.; Kodadek, T.; Burley, S. K. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.* **1999**, *13*, 3217-3230.
(194) Bouvignies, G.; Vallurupalli, P.; Hansen, D. F.; Correia, B. E.; Lange, O.; Bah, A.; Vernon, R. M.; Dahlquist, F. W.; Baker, D.; Kay, L. E. Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature* **2011**, *477*, 111-U134.
(195) Hermann, P.; Fredericq, E. Role of AT Pairs in Acid Denaturation of DNA. *Nucleic Acids Res.* **1977**, *4*, 2939-2947.
(196) Siegfried, N. A.; O'Hare, B.; Bevilacqua, P. C. Driving forces for nucleic acid pK(a) shifting in an A(+).C wobble: effects of helix position, temperature, and ionic strength. *Biochemistry* **2010**, *49*, 3225-3236.
(197) Strobel, S. A.; Cochrane, J. C. RNA catalysis: ribozymes, ribosomes, and riboswitches. *Curr. Opin. Chem. Biol.* **2007**, *11*, 636-643.
(198) Scott, W. G. Ribozymes. *Curr. Opin. Struct. Biol.* **2007**, *17*, 280-286.
(199) Doudna, J. A.; Lorsch, J. R. Ribozyme catalysis: not different, just worse. *Nat. Struct. Mol. Biol.* **2005**, *12*, 395-402.
(200) Ferre-D'Amare, A. R.; Zhou, K. H.; Doudna, J. A. Crystal structure of a hepatitis delta virus ribozyme. *Nature* **1998**, *395*, 567-574.
(201) Chen, J. H.; Yajima, R.; Chadalavada, D. M.; Chase, E.; Bevilacqua, P. C.; Golden, B. L. A 1.9 angstrom Crystal Structure of the HDV Ribozyme Precleavage Suggests both Lewis Acid and General Acid Mechanisms Contribute to Phosphodiester Cleavage. *Biochemistry* **2010**, *49*, 6508-6518.
(202) Rupert, P. B.; Massey, A. P.; Sigurdsson, S. T.; Ferre-D'Amare, A. R. Transition state stabilization by a catalytic RNA. *Science* **2002**, *298*, 1421-1424.
(203) Rupert, P. B.; Ferre-D'Amare, A. R. Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature* **2001**, *410*, 780-786.
(204) Pley, H. W.; Flaherty, K. M.; McKay, D. B. Three-dimensional structure of a hammerhead ribozyme. *Nature* **1994**, *372*, 68-74.
(205) Hoffmann, B.; Mitchell, G. T.; Gendron, P.; Major, F.; Andersen, A. A.; Collins, R. A.; Legault, P. NMR structure of the active conformation of the Varkud satellite ribozyme cleavage site. *Proc. Natl. Acad. Sci. U S A* **2003**, *100*, 7003-7008.
(206) Bevilacqua, P. C. Mechanistic considerations for general acid-base catalysis by RNA: revisiting the mechanism of the hairpin ribozyme. *Biochemistry* **2003**, *42*, 2259-2265.
(207) Wilson, T. J.; Ouellet, J.; Zhao, Z. Y.; Harusawa, S.; Araki, L.; Kurihara, T.; Lilley, D. M. Nucleobase catalysis in the hairpin ribozyme. *RNA* **2006**, *12*, 980-987.
(208) Spitale, R. C.; Volpini, R.; Heller, M. G.; Krucinska, J.; Cristalli, G.; Wedekind, J. E. Identification of an Imino Group Indispensable for Cleavage by a Small Ribozyme. *J. Am. Chem. Soc.* **2009**, *131*, 6093-6095.
(209) Kath-Schorr, S.; Wilson, T. J.; Li, N. S.; Lu, J.; Piccirilli, J. A.; Lilley, D. M. J. General Acid-Base Catalysis Mediated by Nucleobases in the Hairpin Ribozyme. *J. Am. Chem. Soc.* **2012**, *134*, 16717-16724.
(210) Hampel, A.; Cowan, J. A. A unique mechanism for RNA catalysis: the role of metal cofactors in hairpin ribozyme cleavage. *Chem. Biol.* **1997**, *4*, 513-517.
(211) Nesbitt, S.; Hegg, L. A.; Fedor, M. J. An unusual pH-independent and metal-ion-independent mechanism for hairpin ribozyme catalysis. *Chem. Biol.* **1997**, *4*, 619-630.

(212) Nam, K.; Gao, J.; York, D. M. Quantum Mechanical/Molecular Mechanical Simulation Study of the Mechanism of Hairpin Ribozyme Catalysis. *J. Am. Chem. Soc.* **2008**, *130*, 4680-4691.
(213) Mlynsky, V.; Banas, P.; Walter, N. G.; Sponer, J.; Otyepka, M. QM/MM studies of hairpin ribozyme self-cleavage suggest the feasibility of multiple competing reaction mechanisms. *J. Phys. Chem. B.* **2011**, *115*, 13911-13924.
(214) Wilcox, J. L.; Bevilacqua, P. C. pK(a) Shifting in Double-Stranded RNA Is Highly Dependent upon Nearest Neighbors and Bulge Positioning. *Biochemistry* **2013**, *52*, 7470-7476.
(215) Huppler, A.; Nikstad, L. J.; Allmann, A. M.; Brow, D. A.; Butcher, S. E. Metal Binding and Base Ionization in the U6 RNA Intramolecular Stem-loop Structure. *Nat. Struct. Mol. Biol.* **2002**, *9*, 431-435.
(216) Blad, H.; Reiter, N. J.; Abildgaard, F.; Markley, J. L.; Butcher, S. E. Dynamics and metal ion binding in the U6 RNA intramolecular stem-loop as analyzed by NMR. *J. Mol. Biol.* **2005**, *353*, 540-555.
(217) Laricheva, E. N.; Arora, K.; Knight, J. L.; Brooks, C. L., III. Deconstructing Activation Events in Rhodopsin. *J. Am. Chem. Soc.* **2013**, *135*, 10906-10909.
(218) Chen, W.; Wallace, J. A.; Yue, Z.; Shen, J. K. Introducing Titratable Water to All-Atom Molecular Dynamics at Constant pH. *Biophys. J.* **2013**, *105*, 15-17.
(219) Kaljurand, I.; Kutt, A.; Soovali, L.; Rodima, T.; Maemets, V.; Leito, I.; Koppel, I. A. Extension of the self-consistent spectrophotometric basicity scale in acetonitrile to a full span of 28 pK(a) units: Unification of different basicity scales. *J. Org. Chem.* **2005**, *70*, 1019-1028.
(220) Kutt, A.; Leito, I.; Kaljurand, I.; Soovali, L.; Vlasov, V. M.; Yagupolskii, L. M.; Koppel, I. A. A comprehensive self-consistent spectrophotometric acidity scale of neutral bronsted acids in acetonitrile. *J. Org. Chem.* **2006**, *71*, 2829-2838.
(221) Goh, G. B.; Garcia-Moreno, B.; Brooks, C. L., III. The High Dielectric Constant of Staphylococcal Nuclease Is Encoded in Its Structural Architecture. *J. Am. Chem. Soc.* **2011**, *133*, 20072-20075.
(222) Simonson, T. What Is the Dielectric Constant of a Protein When Its Backbone Is Fixed? *J. Chem. Theory Comput.* **2013**, *9*, 4603-4608.
(223) Isom, D. G.; Castaneda, C. A.; Velu, P. D.; Garcia-Moreno, B. Charges in the hydrophobic interior of proteins. *Proc. Natl. Acad. Sci. U S A* **2010**, *107*, 16096-16100.
(224) Kato, M.; Warshel, A. Using a charging coordinate in studies of ionization induced partial unfolding. *J. Phys. Chem. B.* **2006**, *110*, 11566-11570.
(225) Ghosh, N.; Cui, Q. pK(a) of residue 66 in Staphylococal nuclease. I. Insights from QM/MM simulations with conventional sampling. *J. Phys. Chem. B.* **2008**, *112*, 8387-8397.
(226) Di Russo, N. V.; Estrin, D. A.; Marti, M. A.; Roitberg, A. E. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pK(a)s: The Case of Nitrophorin 4. *PLoS Comput. Biol.* **2012**, *8*.
(227) Damjanovic, A.; Brooks, B. R.; Garcia-Moreno, E. B. Conformational Relaxation and Water Penetration Coupled to Ionization of Internal Groups in Proteins. *J. Phys. Chem. A.* **2011**, *115*, 4042-4053.
(228) Chimenti, M. S.; Khangulov, V. S.; Robinson, A. C.; Heroux, A.; Majumdar, A.; Schlessman, J. L.; Garcia-Moreno, B. Structural Reorganization Triggered by Charging of Lys Residues in the Hydrophobic Interior of a Protein. *Structure* **2012**, *20*, 1071-1085.
(229) Goh, G. B.; Eike, D. M.; Murch, B. P.; Brooks, C. L., III. Accurate Modeling of Ionic Surfactants at High Concentration. *J. Phys. Chem. B.* **2015**, *119*, 6217-6224.

(230) Missel, P. J.; Mazer, N. A.; Benedek, G. B.; Young, C. Y.; Carey, M. C. Thermodynamic analysis of the growth of sodium dodecyl sulfate micelles. *J. Phys. Chem.* **1980**, *84*, 1044-1057.
(231) Bezzobotnov, V. Y.; Borbély, S.; Cser, L.; Faragó, B.; Gladkih, I. A.; Ostanevich, Y. M.; Vass, S. Temperature and concentration dependence of properties of sodium dodecyl sulfate micelles determined from small-angle neutron scattering experiments. *J. Phys. Chem.* **1988**, *92*, 5738-5743.
(232) Collura, J. S.; Harrison, D. E.; Richards, C. J.; Kole, T. K.; Fisch, M. R. The effects of concentration, pressure, and temperature on the diffusion coefficient and correlation length of SDS micelles. *J. Phys. Chem. B.* **2002**, *105*, 4846-4852.
(233) Tang, X.; Koenig, P. H.; Larson, R. G. Molecular dynamics simulations of sodium dodecyl sulfate micelles in water - The effect of the force field. *J. Phys. Chem. B.* **2014**, *118*, 3864-3880.
(234) Chen, A. A.; Pappu, R. V. Parameters of monovalent ions in the AMBER-99 forcefield: Assessment of inaccuracies and proposed improvements. *J. Phys. Chem. B.* **2007**, *111*, 11884-11887.
(235) Reif, M. M.; Winger, M.; Oostenbrink, C. Testing of the GROMOS force-field parameter set 54A8: Structural properties of electrolyte solutions, lipid bilayers, and proteins. *J. Chem. Theory Comput.* **2013**, *9*, 1247-1264.
(236) Bruce, C. D.; Berkowitz, M. L.; Perera, L.; Forbes, M. D. E. Molecular dynamics simulation of sodium dodecyl sulfate micelle in water: Micellar structural characteristics and counterion distribution. *J. Phys. Chem. B.* **2002**, *106*, 3788-3793.
(237) Rakitin, A. R.; Pack, G. R. Molecular dynamics simulations of ionic interactions with dodecyl sulfate micelles. *J. Phys. Chem. B.* **2004**, *108*, 2712-2716.
(238) Yoshii, N.; Okazaki, S. A molecular dynamics study of structural stability of spherical SDS micelle as a function of its size. *Chem. Phys. Lett.* **2006**, *425*, 58-61.
(239) Sammalkorpi, M.; Karttunen, M.; Haataja, M. Structural properties of ionic detergent aggregates: A large-scale molecular dynamics study of sodium dodecyl sulfate. *J. Phys. Chem. B.* **2007**, *111*, 11722-11733.
(240) Åqvist, J. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.* **1990**, *94*, 8021-8024.
(241) Dang, L. X. Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: A molecular dynamics study. *J. Am. Chem. Soc.* **1995**, *117*, 6954-6960.
(242) Jensen, K. P.; Jorgensen, W. L. Halide, ammonium, and alkali metal ion parameters for modeling aqueous solutions. *J. Chem. Theory Comput.* **2006**, *2*, 1499-1509.
(243) Joung, I. S.; Cheatham Iii, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B.* **2008**, *112*, 9020-9041.
(244) Weerasinghe, S.; Smith, P. E. A Kirkwood-Buff derived force field for sodium chloride in water. *J. Chem. Phys.* **2003**, *119*, 11342-11349.
(245) Gee, M. B.; Cox, N. R.; Jiao, Y.; Bentenitis, N.; Weerasinghe, S.; Smith, P. E. A Kirkwood-Buff derived force field for aqueous alkali halides. *J. Chem. Theory Comput.* **2011**, *7*, 1369-1380.
(246) Hess, B.; Van Der Vegt, N. F. A. Cation specific binding with protein surface charges. *Proc. Natl. Acad. Sci. U S A* **2009**, *106*, 13296-13300.
(247) Luo, Y.; Roux, B. Simulation of osmotic pressure in concentrated aqueous salt solutions. *J. Phys. Chem. Lett.* **2010**, *1*, 183-189.

(248) Yoo, J.; Aksimentiev, A. Improved parametrization of Li +, Na +, K +, and Mg 2+ ions for all-atom molecular dynamics simulations of nucleic acid systems. *J. Phys. Chem. Lett.* **2012**, *3*, 45-50.
(249) Venable, R. M.; Luo, Y.; Gawrisch, K.; Roux, B.; Pastor, R. W. Simulations of anionic lipid membranes: Development of interaction-specific ion parameters and validation using NMR data. *J. Phys. Chem. B.* **2013**, *117*, 10183-10192.
(250) Jusufi, A.; Hynninen, A. P.; Panagiotopoulos, A. Z. Implicit solvent models for micellization of ionic surfactants. *J. Phys. Chem. B.* **2008**, *112*, 13783-13792.
(251) Gampe, T.; Libuś, Z. Osmotic and activity coefficients of CH3SO4Na(aq) and CH3SO4K(aq) at 25°C. *J. Sol. Chem.* **1999**, *28*, 837-847.
(252) Riske, K. A.; Politi, M. J.; Reed, W. F.; Lamy-Freund, M. T. Temperature and ionic strength dependent light scattering of DMPG dispersions. *Chem. Phys. Lipids.* **1997**, *89*, 31-44.
(253) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656-1676.
(254) Schuler, L. D.; Walde, P.; Luisi, P. L.; van Gunsteren, W. F. Molecular dynamics simulation of n-dodecyl phosphate aggregate structures. *Eur Biophys J Biophy* **2001**, *30*, 330-343.
(255) Elcock, A. H. Atomic-level observation of macromolecular crowding effects: Escape of a protein from the GroEL cage. *Proc. Natl. Acad. Sci. U S A* **2003**, *100*, 2340-2344.
(256) Qin, S.; Zhou, H. X. Atomistic modeling of macromolecular crowding predicts modest increases in protein folding and binding stability. *Biophys. J.* **2009**, *97*, 12-19.
(257) Harada, R.; Sugita, Y.; Feig, M. Protein crowding affects hydration structure and dynamics. *J. Am. Chem. Soc.* **2012**, *134*, 4842-4849.
(258) Dennis, S.; Kortvelyesi, T.; Vajda, S. Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl. Acad. Sci. U S A* **2002**, *99*, 4290-4295.
(259) Guvench, O.; MacKerell Jr, A. D. Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Comput. Biol.* **2009**, *5*.
(260) Lexa, K. W.; Carlson, H. A. Full protein flexibility is essential for proper hot-spot mapping. *J. Am. Chem. Soc.* **2011**, *133*, 200-202.
(261) Lexa, K. W.; Goh, G. B.; Carlson, H. A. Parameter choice matters: Validating probe parameters for use in mixed-solvent simulations. *J. Chem. Inf. Model.* **2014**, *54*, 2190-2199.
(262) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98-105.
(263) Ploetz, E. A.; Smith, P. E. A Kirkwood-Buff force field for the aromatic amino acids. *Phys. Chem. Chem. Phys.* **2011**, *13*, 18154-18167.
(264) Morrow, B. H.; Koenig, P. H.; Shen, J. K. Atomistic simulations of pH-dependent self-assembly of micelle and bilayer from fatty acids. *J. Chem. Phys.* **2012**, *137*.
(265) Morrow, B. H.; Koenig, P. H.; Shen, J. K. Self-assembly and bilayer-micelle transition of fatty acids studied by replica-exchange constant ph molecular dynamics. *Langmuir* **2013**, *29*, 14823-14830.