Constant-pH Simulations with the Polarizable Atomic Multipole AMOEBA Force Field

Andrew C. Thiel, Matthew J. Speranza, Sanika Jadhav, Lewis L. Stevens, Daniel K. Unruh, Pengyu Ren, Jay W. Ponder, Jana Shen, and Michael J. Schnieders*

ABSTRACT: Accurately predicting protein behavior across diverse pH environments remains a significant challenge in biomolecular simulations. Existing constant-pH molecular dynamics (CpHMD) algorithms are limited to fixed-charge force fields, hindering their application to biomolecular systems described by permanent atomic multipoles or induced dipoles. This work overcomes these limitations by introducing the first polarizable CpHMD algorithm in the context of the Atomic Multipole Optimized Energetics for Biomolecular Applications (AMOEBA) force field. Additionally, our implementation in the open-source Force Field X (FFX) software has the unique ability to handle titration state changes for crystalline systems including flexible support for all 230 space groups. The evaluation of constant-pH molecular dynamics (CpHMD) with the AMOEBA force field was performed on 11 crystalline peptide systems that span the titrating amino acids (Asp, Glu, His, Lys, and Cys). Titration states were correctly predicted for 15 out of the 16 amino acids present in the 11 systems, including for the coordination of Zn$^{2+}$ by cysteines. The lone exception was for a HIS-ALA peptide where CpHMD predicted both neutral histidine tautomers to be equally populated, whereas the experimental model did not consider multiple conformers and diffraction data are unavailable for refinement. This work demonstrates the promise polarizable CpHMD simulations for $pK_a$ predictions, the study of biochemical mechanisms such as the catalytic triad of proteases, and for improved protein−ligand binding affinity accuracy in the context of pharmaceutical lead optimization.

INTRODUCTION

A ubiquitous regulator of protein structure is pH, which is exemplified by the impact of low pH on the stability and folding of proteins, even causing proteins to denature. Misregulated pH can contribute to detrimental effects such as the formation of amyloid fibrils in Alzheimer’s disease and insulin aggregation. These effects on structure propagate to protein functions that drive diverse activities such as structural dynamics, ligand binding, and enzyme activity. For example, the M2 protein from the influenza A virus is a pH-activated proton channel that mediates the acidification of the interior of viral particles entrapped in endosomes. Other examples include the regulation of water uptake via aquaporin gating by cytosolic pH in plant root cells and a low-pH induced conformational change in hemagglutinin that activates fusion activity in the influenza virus. Understanding these effects in ligand binding is also critical to drug design where approximately 78% of oral drugs have varying protonation states. The electrostatic contribution to these processes cannot be fully described without accurate knowledge of all fluctuating protonation states, especially considering that for an average protein about 25% of residues have titratable protons and that charge−charge interactions produce some of the strongest atomic forces at work in proteins.

There exist many strategies for the calculation of protein titration curves and residue $pK_a$ values. These techniques vary in both complexity and speed, from rigorous quantum mechanics to empirical rule-based heuristics. Even among those methods grounded in ab initio chemical theory there is a great deal of variety with respect to the treatment of dielectric screening, the calculation of electrostatic contributions, polarization effects, titration coupling, and conformational sampling. Constant-pH molecular dynamics extends past mere $pK_a$ prediction, offering an understanding of titration-coupled conformational dynamics through which many protein functions act.
FOUNDATIONS OF CONSTANT-PH MOLECULAR DYNAMICS

The foundations for most contemporary advanced titration models were described in 1997 by Baptista et al. Recognizing the sensitivity of protonation energies to small changes in atomic coordinates, they devised a method by which protein titration could be observed as a function of these movements (i.e., during the course of a molecular dynamics simulation). Their technique, known as constant-pH molecular dynamics (CpHMD), combines molecular mechanics with continuum electrostatic pH calculations by adding the charge state at each titratable site as a parameter in the MD simulation. Fractional protonation states are made available to each site and are updated after a defined number of MD steps via Poisson–Boltzmann (PB) calculations.

Five years later, Baptista et al. modified their CpHMD method to affect a departure from unphysical fractional protons. Rather than allow continuously variable charge on titratable residues, their solution employed a Metropolis Monte Carlo (MC) criterion at regular intervals during simulation to switch directly and instantaneously between fully protonated and fully deprotonated states. The PB equation was used in evaluating the MC moves between the titration states. Mongan, Case, and McCammon employed the generalized Born (GB) equation in their CpHMD methods to improve the sampling speed over PB methods. Birgi and colleagues introduced a variation on Baptista’s theme where a costly explicit-solvent thermodynamic free energy simulation was used to estimate the free energy of deprotonation for each MC move. A significant acceleration to sampling discrete protonation states was achieved by Roitberg and colleagues through their implementation of replica exchange on hybrid MD/MC simulations. Although nonphysical intermediate states are not sampled using discrete approaches, much of the recent work in this area has continued to use a continuous approach, as it can easily be applied to all-atom simulations and is efficient. However, nonequilibrium MD/MC approaches are being explored to overcome the issue of the protonation state flip being rejected in MC. One continuous approach pioneered by Brooks and co-workers was the application of λ-dynamics to update titration states. In this approach, a λ-coordinate corresponding to each titration state is added to create an extended Hamiltonian. This λ-coordinate is propagated according to the equations of motion governed by the potential of its surroundings and the environmental pH. The λ-dynamics on titratable sites approach has become more widely adopted and can now be found within the GROMACS, CHARMM, and Amber MD packages. Continuous approaches have been expanded to optionally include long-range electrostatics via particle-mesh Ewald (PME) summation by Shen et al. and titratable waters to maintain zero net charge for PME.

EFFICIENT ALCHEMICAL PATHS IN A POLARIZABLE MULTIPOLe CONTEXT

At present, the state-of-the-art implementation of arbitrary thermodynamic path sampling under a polarizable multipole potential is to compute the electrostatic energy of a dual topology configuration via linear interpolation between end states. The self-consistent fields (SCFs) for each topology are converged separately prior to interpolation. Since convergence of the SCF is a bottleneck in the evaluation of the energy and forces for a polarizable multipole model, the need to compute two SCFs makes dual topology calculations relatively less attractive. Here we describe, for the first time, an analytical thermodynamic path for AMOEBA that requires only a single SCF evaluation per energy and gradient evaluation.

THEORY AND IMPLEMENTATION

Continuous CpHMD for the Polarizable AMOEBA Model with PME. The current implementation supports the residues ASP, GLU, HIS, LYS, and CYS. Following a common approach established by fixed partial charge CpHMD models, the protonation states of titratable residues are propagated alongside the atomic Cartesian coordinates during dynamics. This is described by an extended Hamiltonian laid out by Shen and co-workers:

\[
H(X, \theta) = U_{\text{bond}}(X) + U_{\text{nonbond}}(X, \theta) + U^\theta(\theta) = \sum_i^N 0.5 m_i X_i^2 + \sum_k^N 0.5 m_k \theta_k^2 \tag{1}
\]

where \(i\) is the index over atoms and \(k\) is the index over both titration and tautomer extended system variables, \(X\) is the Cartesian coordinate vector, and \(\theta\) is a vector of both titration and tautomer extended system states. The titration (\(\lambda_i\)) and tautomer (\(\zeta_k\)) states are bound between 0 and 1 through the relation \(\lambda_i = \sin^2(\theta_i)\). The first term on the right-hand side of eq 1 is the potential due to all bonded terms. The bonded terms were not scaled with titration states such that a deprotonated atomic site retains bonded terms with an alchemically decoupled proton. The second term is the nonbonded potential, which is a function of both Cartesian and titration/tautomer coordinates. The nonbonded energy terms and their derivatives are described in the next section. The third term is the biasing potential, which is a function of only the titration and tautomer states. The bias is composed of three components:

\[
U^\theta(\theta) = -U_{\text{barr}}(\theta) - U_{\text{mod}}(\theta) + U_{\text{PMF}}(\theta) \tag{2}
\]

The first component represents a barrier centered at 0.5 that suppresses the intermediate titration and tautomer states.

\[
U_{\text{barr}}(\lambda_k, \zeta_k) = 4\beta \left( \lambda_k - \frac{1}{2} \right)^2 + 4\beta \left( \zeta_k - \frac{1}{2} \right)^2 \tag{3}
\]

This barrier also affects the kinetics of the transition between end states by specifying the magnitude of the barrier (\(\beta\)). The \(\zeta\)-dependent term is not present for LYS and CYS residues that lack tautomers. The model potential, \(U_{\text{mod}}(\theta)\), is a potential of mean force (PMF) for protonation of the model compound

\[
U_{\text{mod}}(\lambda_k) = A\lambda_k^3 + B\lambda_k^2 + C\lambda_k \tag{4}
\]

This PMF flattens the free-energy surface between the protonated and deprotonated states such that the model compound’s titration state is dictated by its environment of nonbonded interactions and the final bias term. The PMF shown in eq 4 applies to all residues where there is no tautomorphism (the PMF functional form for other cases is given in the Supporting Information). The symmetry of ASP and GLU oxygen atoms implies that their PMF will show no energy difference between end states. In the case of HIS, three PMFs were collected. Two for the protonation between the distinct tautomers, HIE and HID, and one between the two
tautomer states (see Simulation Protocol). The pH bias, \( U_{ph}(\theta) \) is the final term and represents the relative free energy of deprotonation at the environmental pH

\[
U_{ph}(\lambda, \zeta) = \ln(10) k_B T (1 - \lambda_{k}) \left[ \zeta_0 \left( pK_a^1 - \text{pH} \right) + (1 - \zeta_0) \left( pK_a^2 - \text{pH} \right) \right]
\]

where \( k_B \) and \( T \) are the Boltzmann constant and temperature, respectively. The scaling by \((1 - \lambda_{k})\) is consistent with the convention used here whereby \( \lambda_{k} = 1 \) denotes a protonated site. As before, when LYS and CYS are considered, there is no \( \zeta_0 \)-dependence.

van der Waals Interactions. Under the AMOEBA model, a buffered 14-7 van der Waals term is used, which can be modified to include soft-core capabilities that allow for smooth decoupling as \( \lambda \) goes to zero. The soft-core form between atoms \( i \) and \( j \) is given by

\[
U_{vdw}(\lambda, r) = (\lambda)^{\delta} e_i f_{ij}(\xi)\lambda
\]

where \( t_1 \) and \( t_2 \) are defined as

\[
t_1 = \frac{(1 + \delta)^{\rho - m}}{(\rho + \delta)^{\rho - m}}
\]

\[
t_2 = \frac{1 + \rho}{\rho^{\rho - m}} - 2
\]

and the buffered 14-7 constants are

\( n = 14, \ m = 7, \ \delta = 0.07, \ \gamma = 0.12 \)

In the above expressions, \( e_i \) is the well depth and \( \rho \) is the normalized atomic separation distance \((r_{ij}/r_{min,q})\) defined by the ratio of the minimum-energy separation distance \( r_{min,q} \) and the current separation distance \( r_{ij} \).

In the case of an alchemical proton, preliminary work demonstrated that the soft-core term is not necessary because each proton is protected by the \( v_d \) of the van der Waals repulsion provided by its heavy atom. For this reason, the extended system van der Waals functional form given by

\[
U_{vdw}(\lambda, \zeta, r) = (f(\lambda_1, \zeta) f(\lambda_2, \zeta))^\gamma U_{vdw}(i, j)
\]

is simply a scaled version of the default (nonsoft core) van der Waals energy, where the scaling is controlled by titration and tautomer variables. These scaling contributions are separable and are defined for each residue in Table 1.

Permanent Atomic Multipoles as a Function of Titration and Tautomer States. The permanent atomic multipole for atom \( i \) is described by a vector of multipole coefficients

\[
M_i = [q_i, d_{x,i}, d_{y,i}, d_{z,i}, \Theta_{a2i}, \Theta_{b2i}, \Theta_{c2i}, \Theta_{a3i}, \Theta_{b3i}, \Theta_{c3i}]
\]

If the multipole is a function of only a titration state variable \( \lambda_{k} \) as is the case with lysine, then \( M_i(\lambda_{k}) \) is given by a linear function of the permanent atomic moments for the unprotonated \( M_i^{(U)} \) and protonated \( M_i^{(P)} \) states

\[
M_i(\lambda_{k}) = (1 - \lambda_{k})M_i^{(U)} + \lambda_{k}M_i^{(P)}
\]

(10)

and the partial derivative of \( M_i(\lambda_{k}) \) with respect to \( \lambda_{k} \) is

\[
\frac{dM_i}{d\lambda_{k}} = M_i^{(P)} - M_i^{(U)}
\]

(11)

If the multipole is also a function of the tautomer state variable \( \zeta_0 \), then \( M_i(\lambda_{k}, \zeta_0) \) is given by a mixing of linear functions between permanent moments for the unprotonated and protonated states and between tautomeric moments. In the case of protonated tautomerism, there is mixing between the unprotonated moment and two protonated moments \( M_i^{(P_1)} \) and \( M_i^{(P_2)} \)

\[
M_i(\lambda_{k}, \zeta_0) = (1 - \lambda_{k})M_i^{(U)} + \lambda_{k}(\zeta_0M_i^{(P_1)} + (1 - \zeta_0)M_i^{(P_2)})
\]

(12)

and the partial derivative of \( M_i(\lambda_{k}, \zeta_0) \) with respect to \( \lambda_{k} \) is

\[
\frac{dM_i}{d\lambda_{k}} = M_i^{(P_2)} - M_i^{(P_1)}
\]

(13)

and the partial derivative with respect to \( \zeta_0 \) is

\[
\frac{dM_i}{d\zeta_0} = \lambda_{k}(M_i^{(P_1)} - M_i^{(P_2)})
\]

(14)

The rotation of multipole moments from their local chemical frame into the global frame is performed for both \( M_i \) and its partial derivatives \( \partial M_i \) and \( \partial^2 M_i \) after the functions defined above have been evaluated. For this reason, both the local coordinate frame convention (e.g., Z-then-X or Z-then-bisector) and its frame-defining atoms must be identical for both unprotonated and protonated end states.

Real Space Permanent Multipole Energy and Partial Derivatives. To describe the permanent multipole energy, we introduce multipolar operators

\[
\hat{L}_i = (q_i + d_i \nabla + \Theta_i : \nabla \nabla)
\]

(15)

\[
\hat{L}_j = (q_j - d_j \nabla + \Theta_j : \nabla \nabla)
\]

(16)

The partial derivative of \( \hat{L}_i \) or \( \hat{L}_j \) with respect to \( \lambda_{k} \) is achieved by substituting the rotated \( \hat{M}_i \) or \( \hat{M}_i \) multipole moments into eq 15 or eq 16, respectively, to give

\[
\frac{\partial\hat{L}_i}{\partial\lambda_{k}} = \hat{L}_i = (q_i + d_i \nabla + \Theta_i : \nabla \nabla)
\]

(17)

\[
\frac{\partial\hat{L}_j}{\partial\lambda_{k}} = \hat{L}_j = (q_j - d_j \nabla + \Theta_j : \nabla \nabla)
\]

(18)

The partial derivative of \( \hat{L}_i \) or \( \hat{L}_j \) with respect to \( \zeta_0 \) is the same as the above but with rotated \( \hat{M}_i \) or \( \hat{M}_i \) from eq 14 substituted instead.

Let \( \mathbf{n} \) be a vector of integer triples such that \( \mathbf{n} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c} \) where \( \{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \) are the unit cell vectors along the \( a, b, \) and \( c \) axes. Then summing over \( \mathbf{n} \) accomplishes the interaction with periodic images. For the case of \( \mathbf{n} = \{0, 0, 0\} \), \( i = j \) is omitted (i.e., self-interactions within the central unit cell, which is denoted by the prime on \( \mathbf{n} \) below).
\[
U_{\text{perm}}(X, \lambda, \zeta) = \frac{1}{2} \sum_{n} \sum_{i} \sum_{j} \hat{T}_{ij} \hat{T}_{ij} \left( \frac{1}{f_{ij}} \right)
\]

(19)

where \( f_{ij} = r_{ij} + \alpha(1 - \lambda \zeta)^{2} \) augments \( r_{ij} = |r_{i} - r_{j} + n\ell | \) with soft-core support.

Using the Ewald range separation approach, the short-range real space interaction potential is given by

\[
\phi^\text{(real)}_{\text{perm}}(X, \lambda, \zeta) = \sum_{n=1}^{N} \sum_{i=1}^{N} \frac{\text{erfc}(\beta f_{ij})}{f_{ij}}
\]

(20)

where \( \beta \) is the tunable Ewald parameter. The real space permanent multipole energy is then given by

\[
U^\text{(real)}_{\text{perm}}(X, \lambda, \zeta) = \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} \frac{\text{erfc}(\beta f_{ij})}{f_{ij}}
\]

(21)

Differentiating the real space permanent energy with respect to \( \lambda_{k} \) gives

\[
\frac{\partial U^\text{(real)}_{\text{perm}}(X, \lambda, \zeta)}{\partial \lambda_{k}} = \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} \frac{\text{erfc}(\beta f_{ij})}{f_{ij}}
\]

\[+ \hat{T}_{ij} \text{erfc}(\beta f_{ij}) f_{ij} + \hat{T}_{ij} \frac{\partial}{\partial \lambda_{k}} \left( \frac{\text{erfc}(\beta f_{ij})}{f_{ij}} \right) \]

(22)

The derivative of the real space permanent energy with respect to \( \zeta_{k} \) is the same as above but uses \( \hat{T}_{ij} \) and \( \hat{T}_{ij} \). Titrating hydrogens can be treated with soft-core electrostatic interactions to protect their charge site at small values of \( \lambda \) when their van der Waals interactions are not present. However, the titrating hydrogen atoms are not treated with soft-core electrostatics for the same reason described above for neglecting a soft core for the van der Waals term. In this case, the third term of eq 22 is eliminated because \( f \) is not a function of \( \lambda \).

Reciprocal Space Permanent Multipole Energy and Partial Derivatives. From Sagui et al., the reciprocal space multipolar electrostatic energy \( U_{\text{rec}} \) is given by

\[
U_{\text{rec}} = \frac{1}{2} \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} Q^{R}(m) \cdot (G^{R} Q^{R})(m)
\]

(23)

where \( Q^{R} \) is the reciprocal lattice grid populated with the splined multipoles \( \theta_{i} \)

\[
Q^{R}(k_{1}, k_{2}, k_{3}) = \sum_{i=1}^{N} \sum_{n} \hat{T}_{ij} \theta_{i} (u_{ii} - k_{1} - n_{1}K_{1})
\]

\[\times \theta_{j}(u_{jj} - k_{2} - n_{2}K_{2})
\]

\[\times \theta_{k}(u_{kk} - k_{3} - n_{3}K_{3})
\]

(24)

And \( G^{R} \) is the discrete Fourier transform of the coefficients arising from the structure factor. The partial derivative of eq 23 with respect to \( \lambda_{k} \) is given as

\[
\frac{\partial U_{\text{rec}}}{\partial \lambda_{k}} = \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} \sum_{s=0}^{K-1} Q^{R}(m) \cdot (G^{R} Q^{R})(m)
\]

\[\frac{\partial}{\partial \lambda_{k}} \]

(25)

In other words, the splined \( M \), (eq 12) is the reciprocal space source term, and its reciprocal space potential is felt by \( \partial M/\partial \lambda_{k} = M_{k} \) or \( \partial M/\partial \zeta_{k} = M_{k} \) (eq 13, eq 14). This is accomplished by computing the reciprocal space potential using \( M \) and then evaluating the partial derivatives using \( M_{k} \) or \( M_{k} \).

Polarizability as a Function of Titration and Tautomer States. The AMOEBA polarization model is based on isotropic polarizabilities and, in general, a single polarizability value per element (i.e., all carbons share the same polarizability value). Polarizabilities that change as a function of a titration state \( \lambda_{k} \) and tautomer state \( \zeta_{k} \) can be defined as \( \alpha(\lambda_{k}, \zeta_{k}) \) using a linear function between unprotonated \( \alpha^{(0)} \) and protonated \( \alpha^{(p)} \) states and a linear function between tautomer states. This is analogous to the functions used to interpolate atomic multipole coefficients. Equation 26 is an example of the case of protonated tautomerism

\[
\alpha(\lambda_{k}, \zeta_{k}) = (1 - \lambda_{k})\alpha^{(U)} + \lambda_{k}\zeta_{k}\alpha^{(P1)} + (1 - \zeta_{k})\alpha^{(P2)}
\]

(26)

For most heavy atoms in the AMOEBA protein force field, polarizability is independent of titration state or tautomeric state, while for titrating hydrogen, the expression reduces to linearly turning the polarizability on (or off). The exception to this rule is for sulfur and carboxylic oxygen atoms, such that interpolation occurs between two nonzero polarizabilities. A complication that arises from interpolating between two nonzero values is in the Thole damping model, where the charge density \( \rho \) takes the form

\[
\rho = \frac{3\alpha}{4\pi} \exp(-au^{2})
\]

(27)

where \( u = R_{ij}/(\alpha\zeta)^{1/6} \) is the effective separation as a function of the two polarizabilities at sites \( i \) and \( j \). As the polarizability is a function of state variables \( (\lambda_{k}, \zeta_{k}) \), so too must be the charge density. However, an approximation made by this model is to fix the charge density such that there is no derivative of the charge density with respect to either \( \lambda_{k} \) or \( \zeta_{k} \). Therefore, titrating sulfur and carboxylic oxygen atom charge densities are fixed to a value as calculated by the average of their end-state polarizabilities. This is not necessary for titrating hydrogen atoms where their polarizability will be zero at one end state and will not contribute in any interaction.

Polarization Energy and Partial Derivatives. The following derivation will be presented in the context of a vacuum calculation for simplification, such that the field contributions exclude reciprocal space and self-correction terms. However, the derivatives of these terms will be given following real space derivation. The polarization energy is given by an inner product

\[
U_{\text{pol}}(X, \lambda, \zeta) = -\frac{1}{2} \mu \cdot E_{p}
\]

(28)

where \( \mu \) and \( E_{p} \) are vectors of length \( 3n \) that contain the induced dipoles and permanent multipole field components, respectively. The subscript \( p \) indicates that 1–2, 1–3, etc. bond dipole masking rules are applied in calculating the field at each induced dipole site. For the AMOEBA model, the induced dipole at site \( i \) is a function of an isotropic polarizability \( \alpha_{i} \) interacting with the total field \( \mu \)

\[
\mu = \alpha E_{\text{total},d} = \alpha(\lambda, \zeta)(T_{2}M + T_{1}^{T}\mu)
\]

(29)
Here, $M$ is a vector of length $13n$ that contains multipole moments. Given the definition of $T_{ij}$

$$T_{ij} = \begin{pmatrix}
1 & \frac{\delta}{\delta x_i} & \frac{\delta}{\delta y_j} & \frac{\delta}{\delta z_j} \\
\frac{\delta}{\delta x_i} & \frac{\delta^2}{\delta x_i \delta y_j} & \frac{\delta^2}{\delta x_i \delta z_j} & \frac{\delta^2}{\delta x_i \delta z_j} \\
\frac{\delta}{\delta y_j} & \frac{\delta^2}{\delta x_i \delta y_j} & \frac{\delta^2}{\delta y_j \delta z_j} & \frac{\delta^2}{\delta y_j \delta z_j} \\
\frac{\delta}{\delta z_j} & \frac{\delta^2}{\delta x_i \delta z_j} & \frac{\delta^2}{\delta y_j \delta z_j} & \frac{\delta^2}{\delta y_j \delta z_j}
\end{pmatrix} \cdots \frac{1}{f_{ji}}$$

as the multipole–multipole interaction matrix between sites $i$ and $j$. $T_{ij}$ is the $3 \times 13$ interaction matrix corresponding to the second through fourth rows of $T_{ik}$. In the condensed phase under periodic boundary conditions, $T_{ij}$ will use erfc $(\beta f_{ij})/f_{ij}$ in place of $1/r_{ij}$. $T_{ij}$ is a super matrix with elements of tensor $T_{ij}$ (varying $i$ and $j$). Subscript $d$ indicates the use of the AMOEBA group-based polarization masking rules used to compute the direct field due to permanent atomic multipoles. In the same vein, $T^{11}$ is a supermatrix with elements corresponding to the tensor $T_{ij}^{11}$, which is a submatrix of $T_{ij}$ that contains only the dipole–dipole interaction terms and in condensed phase would include the Thole damping modification to elements of the matrix. We solve for the induced dipoles and then factor out $\mu$:

$$\alpha^{-1} \mu - T^{11} \mu = T^{11} M$$

$$\langle \alpha^{-1} - T^{11} \rangle \mu = T^{11} M$$

(31) (32)

For convenience, a quantity $C$ is defined and substituted above to simplify

$$C = \alpha^{-1} - T^{11}$$

$$\mu = C^{-1} E_d$$

(33) (34)

where $E_d = T^{11} M$ is the permanent field using polarization group masking rules. The partial derivative of the polarization energy with respect to an atomic coordinate is then given by the product rule as

$$\frac{\partial U_{pol}(X, \lambda, \zeta)}{\partial x_i} = -\frac{1}{2} \left[ \frac{\partial E_d}{\partial x_i} C^{-1} E + E_d \frac{\partial C^{-1}}{\partial x_i} + E_d C^{-1} \frac{\partial E_d}{\partial x_i} \right]$$

(35)

Additional simplification is achieved through the use of the following two relationships:

$$\nu = C^{-1} E_p$$

$$\frac{\partial C^{-1}}{\partial x_i} = C^{-1} \frac{\partial C}{\partial x_i} C^{-1} = -C^{-1} \frac{\partial T^{11}}{\partial x_i} C^{-1}$$

(36) (37)

This gives the partial derivative of polarization energy with respect to an atomic coordinate of site $i$:

$$\frac{\partial U_{pol}(X, \lambda, \zeta)}{\partial x_i} = -\frac{1}{2} \left[ \frac{\partial E_d}{\partial x_i} \nu - \mu \frac{\partial T^{11}}{\partial x_i} \nu + \mu \frac{\partial E_d}{\partial x_i} \right]$$

(38)

The derivation with respect to an extended system variable proceeds in a similar fashion:

$$\frac{\partial U_{pol}(X, \lambda, \zeta)}{\partial \lambda_k} = -\frac{1}{2} \left[ \frac{\partial M^{T_d}}{\partial \lambda_k} C^{-1} E_p + E_d \frac{\partial C^{-1}}{\partial \lambda_k} E_p + E_d C^{-1} \frac{\partial M^{T_d}}{\partial \lambda_k} \right]$$

(39)

Using the expression for the lambda derivative of the quantity $C$ gives

$$\frac{\partial C^{-1}}{\partial \lambda_k} = -C^{-1} \frac{\partial C}{\partial \lambda_k} C^{-1} = -C^{-1} \frac{\partial \alpha}{\partial \lambda_k} C^{-1} = C^{-1} \frac{\partial \alpha}{\partial \lambda_k} C^{-1}$$

(40)

Substituting into the central term on the right-hand side of eq 39 gives

$$\frac{\partial U_{pol}(X, \lambda, \zeta)}{\partial \lambda_k} = -\frac{1}{2} \left[ \frac{\partial M^{T_d}}{\partial \lambda_k} C^{-1} E_p + E_d C^{-1} \frac{\partial T^{11}_d}{\partial \lambda_k} + \frac{\partial \alpha}{\partial \lambda_k} \frac{\partial T^{11}_d}{\partial \lambda_k} \right]$$

(41)

Rearrangement of eq 29 using eq 34 reveals the following relationships

$$E_{total,p} = \alpha^{-1} \nu = \alpha^{-1} C^{-1} E_p$$

$$E_{total,d} = \alpha^{-1} \mu = \alpha^{-1} C^{-1} E_d$$

(42) (43)

This allows the simplification of eq 41 to

$$\frac{\partial U_{pol}(X, \lambda, \zeta)}{\partial \lambda_k} = -\frac{1}{2} \left[ \frac{\partial M^{T_d}}{\partial \lambda_k} C^{-1} E_p + \frac{\partial T^{11}_d}{\partial \lambda_k} + \frac{\partial \alpha}{\partial \lambda_k} \frac{\partial T^{11}_d}{\partial \lambda_k} + \mu \frac{\partial T^{11}_d}{\partial \lambda_k} \right]$$

(44)

The first and third terms are equivalent in form to the computation of a polarization energy, as defined in eq 28, but using partial derivatives of the permanent multipole moments with respect to a titration or tautomer variable. The central term is trivial to compute because the $3n \times 3n$ matrix $\alpha$ is diagonal and the only elements of $\partial \alpha/\partial \lambda_k$ that are nonzero arise from polarizabilities that are a function of $\lambda_k$ (e.g., the titrating hydrogen, sulfur, or carboxylic oxygen atoms of a single residue).

In eq 44, the total field is due to permanent multipoles and induced dipoles. The permanent field can be broken down further into real, reciprocal, and self-contributions

$$E_d = E_{real,d} + E_{recip} + E_{self,d}$$

(45)

$$E_p = E_{real,p} + E_{recip} + E_{self,p}$$

(46)

While the middle term on the right-hand side of eq 44 explicitly uses the total field, the first and last terms use the permanent multipole field. These terms resemble the interaction of the titrating multipoles with the field of the induced dipoles. Now considering these terms in the condensed phase with a reciprocal space component and reorganizing eq 44, we can simply define the derivatives as
\[ \frac{\partial U_{\text{pol,recip}}(X, \lambda, \zeta)}{\partial \lambda} = -M^T E_{\text{recip,dipole}} \]  
where \( E_{\text{recip,dipole}} = (E_{\text{recip,\mu}} + E_{\text{recip,\nu}})/2 \). The same logic can be followed for the self-correction term,

\[ \frac{\partial U_{\text{pol,self}}(X, \lambda, \zeta)}{\partial \lambda} = -\mu^T E_{\text{self,dipole}} \]

where \( E_{\text{self,dipole}} = \frac{2e^2}{\lambda \epsilon} (\nu + \mu) \).

**Modification of AMOEBA Force Field Parameters.** As discussed previously,\(^{40}\) dipoles and quadrupoles are defined in a local frame based on the coordinates of neighboring atoms. A crucial element for the interpolated multipole is that their frames and frame-defining atoms must be consistent between the end states. Otherwise, if a frame (e.g., Z-then-X or Z-then-bisector) or its frame-defining atoms at end states differ, then the torque contributions need to be collected separately (e.g., the multipolar contributions from HIS, HID, and HIE end-states would need to be rotated into the global frame separately and their respective torques accumulated separately). Therefore, all permanent atomic multipoles impacted by titration were made consistent with their counterpart(s). Fortunately, this affected only a few titration sites for the carboxylic acids and cysteine (see the updated AMOEBA protein force field parameters in the Supporting Information). Analogous to permanent atomic multipole frame definitions that require consistency, polarization groups for each atomic site must also concur. The permanent multipoles of atoms that belong to the same polarization group do not contribute to the field that induces the dipoles of other atoms within their group. If there is a difference in how the polarization groups are defined between titration end states, then the direct field cannot be defined in a way that is consistent with the use of a single self-consistent field calculation. Fortunately, the only polarization group that required modification was for the deprotonated cysteine.

Finally, an additional small approximation to simplify the algorithm was to consider the bonded terms invariant to titration variables, which was also made in previous work.\(^{28,29,34}\) This was achieved by fixing the bonded terms to those of the protonated state. In the case of carboxylic acids, force field terms were defined as if both oxygen atoms were bonded to a “dummy” hydrogen. Over the course of a simulation, the carboxylic acid could have only one hydrogen atom present with full-strength nonbonded interactions; however, both oxygen atoms always maintain a hydrogen via standard bonded terms.

**pH Replica Exchange.** Due to the computational expense of CphMMD simulations in the AMOEBA force field, a pH-based replica exchange (pH-REX) protocol\(^{44}\) was implemented to accelerate convergence through enhanced sampling. The protocol involves running multiple simulations at different pH values simultaneously and periodically exchanging pH’s between simulations according to a probability given by the Metropolis criterion

\[ P = \begin{cases} 
1 & \text{if } \Delta E \leq 0 \\
\exp(-\beta \Delta E) & \text{otherwise}
\end{cases} \]

where \( \beta \) is given by \( 1/(k_B T) \) and \( \Delta E \) is defined as

\[ \Delta E = U_{\text{pH}}(X_{A \rightarrow B^+} \theta_A \rightarrow B) + U_{\text{pH}}(X_{B \rightarrow A^+} \theta_B \rightarrow A) - (U_{\text{pH}}(X_{A} \theta_A) + U_{\text{pH}}(X_{B} \theta_B)) \]

where A and B represent the two ensembles considered in the exchange.

The force exerted on a proton in response to a new pH that occurs after an exchange often encourages the titration-tautomer coordinates to overcome energetic barriers. This enhances the sampling by allowing for more frequent barrier crossing than would be possible in a simulation with a single fixed pH. This protocol also lends itself well to the generation of titration curves, where each pH window in the replica exchange contributes to the titration curve. For the predictions of the model and protein pK\(_{a}\) values, an exchange was attempted every 500 molecular dynamics steps.

**SIMULATION PROTOCOL**

**Crystal Peptide System Preparation.** All systems were set up and simulated in the open-source Force Field X (FFX) software package. Ten previously studied crystal peptides were obtained from the Cambridge Structural Database\(^{45}\) in CIF format. Their database IDs are as follows: RAVZAQ, RAVZEU, RAVZIY, RAVZOE, RAVZUK, RAWBAT,\(^{46}\) JUKMOR,\(^{47}\) TEKNAY,\(^{48}\) CURLOQ,\(^{49}\) and CUFFUG.\(^{50}\) Each structure had protons added to each titratable site, and a P1 system was prepared in addition to the asymmetric unit (excluding CURLOQ due to reasons mentioned below). Conversion from CIF format to Tinker XYZ format\(^{51}\) with atom types was facilitated by the FFX command ImportCIF.

**Crystal Peptide CphMMD Production Simulations.** The asymmetric units and unit cells were set up according to the deposited crystal records through FFX’s capability to consistently handle symmetry for all space groups with PME.\(^{43}\) The asymmetric system experienced the same crystal packing as the P1 system except for having fewer independently titrating sites. Langevin dynamics simulations, including fluctuating protonation states, were completed at a neutral pH of 7.0 for 10 ns in the NVT ensemble. The temperature was set to 193 K to match the conditions of the previously studied crystals.\(^{46}\) Each of the titratable sites was started at an intermediate state (\( \lambda = 0.5, \zeta = 0.5 \)) to facilitate the titration state quickly sampling an equilibrium value. PME was used for electrostatics with a cutoff of 7 Å, an Ewald alpha of 0.545, and a neutrality constraint.\(^{52}\) For van der Waals interactions, a switching function was used to smoothly switch off interactions over a window of 10 to 12 Å.

**Peptide Model System Preparation.** Capped pentapeptides were used as the model compounds for the determination of the model bias for each residue (eq 4) and took the form of Ace-AAXA-Nme (X = ASP, GLU, HIS, LYS, CYS). The model peptide compounds were solvated in a 40 Å box of explicit water, where the padding of water was at least 10 Å. Finally, the solvated systems were minimized using the L-BFGS algorithm.\(^{53}\) All atoms were represented using the AMOEBA-BIO-2018 force field\(^{21}\) with the modified parameters discussed above and given in the Supporting Information.

**Model Parameterization.** Model parametrization began by collecting the potential of mean force (PMF) curves for the titration of each model compound. This was accomplished by running 50 ns of Langevin dynamics for each of 11 evenly spaced titration windows spanning \( \lambda = 0 \) (deprotonated) to \( \lambda = 1 \) (protonated). The conditions for the dynamics were NVT at 193 K and a temperature step size of 0.25 K. A bias field of 250 kJ/mol/Å was used to pull the system towards the desired titration state. A 100 ns simulation was performed for each condition. The PMF was calculated using the free energy perturbation (FEP) technique with the absolute value of the PMF being

\[ \Delta G = \sum \Delta G_i \]

where \( \Delta G_i \) is the free energy change for each of the 11 steps in the titration. The PMF was then used to determine the model parameters for each compound.
298 K. In each of these windows, the \( \lambda \) value is fixed so that only conformational sampling is collected (i.e., the titration/tautomer state does not fluctuate). The Bennet acceptance ratio method was then used to estimate the free-energy difference between neighboring windows, and the resulting PMF was fit to the polynomial model bias equation (eq 4). As mentioned above, HIS required three such PMFs to be generated. Two titration PMFs were generated at fixed tautomer values of 0 and 1 to collect the PMF of protonating HID and HIE, respectively. A third PMF was generated to shift the proton from \( \delta \)-N to \( \epsilon \)-N by running dynamics over tautomer windows (rather than titration). Once these PMFs were in place as model bias terms, they were validated against three replicate pH-replica exchange molecular dynamics simulations. Titration curves for the model compounds were generated, and the predicted \( pK_a \) for the models was assessed.

The titration curve was generated from a pH range centered on the \( pK_a \) of the model and separated by 0.5 pH unit by using 10 ns simulations. The Langevin dynamics for each replica were carried out under NVT conditions using a 1 fs time step at 298 K. PME was used for electrostatics with a cutoff of 7 Å. Lennard-Jones interactions were switched off over the range of 10–12 Å. Although the forces due to the titration or tautomer variables cannot yet be calculated on a GPU, a hybrid scheme has been adopted to achieve some acceleration on the protein simulations. This scheme involves cycling MD steps between the CPU and GPU. On the CPU, the full sets of forces can be calculated, so the titration states and conformational states can be updated simultaneously. Once a desired number of steps has been reached, the CPU can send the interpolated multipoles and polarizabilities to OpenMM to perform conformational dynamics on these “frozen” titration states. No changes to OpenMM were required because interpolated multipole and polarizability values were handled analogously to the standard AMOEBA force field parameters for amino acid end states. This permits continuous titration states while avoiding the need for Metropolis Monte Carlo moves between discrete end states. After OpenMM completed a specified number of steps, the CPU continues from the new coordinates to begin the next cycle (see the Supporting Information for a detailed description). Although full acceleration through GPUs is a future goal, this scheme has allowed more efficient sampling convergence than CPU moves alone without requiring the OpenMM AMOEBA plugin code to be modified.

\[ S^{\text{depot}} = \frac{1}{1 + 10^{\left(pK_a - \text{pH}\right)}} \]  

**Figure 1.** Titration curves are generated for each pentapeptide. Listed for each are the predicted \( pK_a \) value, the experimental value, and the Hill coefficient.
where $pK_a$ and $n$ are fit parameters and $n$ represents the Hill coefficient. The deprotonated fraction can be defined by simulation counts as

$$S_{\text{deprot}} = \frac{N_{\text{deprot}}}{(N_{\text{deprot}} + N_{\text{prot}})},$$

where $N$ is the number of counts for the respective deprotonated or protonated state.

**Experimental Protocol.** A crystal structure was generated for this work of an $L$-histidyl-$L$-alanine (HIS-ALA) dipeptide in an attempt to further analyze a structure from a previous study. The compound was purchased from Ambeed, Inc. and was received as a salt with trifluoroacetic acid (TFA) instead of the desired pure form. It was crystallized by the vapor diffusion of ethanol into an aqueous solution of dipeptide using the hanging drop method. The experiment was set up in a 24-well plate wherein 1000 $\mu$L of ethanol was added to each well and 2 $\mu$L of concentrated aqueous solution of dipeptide was placed on the coverslip. The coverslip was placed in an inverted position on the well, and the walls of the well were sealed with grease. The plate was kept in a vibration-free zone until single crystals suitable for XRD were visible.

**X-ray Crystallography.** Data were collected on a Bruker D8 VENTURE DUO diffractometer equipped with an IµS 3.0 microfocus source operated at 55 W (50 kV, 1.1 mA) to generate Cu Kα radiation ($\lambda = 1.54178$ Å) and a PHOTON III detector. The crystal was maintained at 100 K throughout the duration of the experiment. Data collection, initial indexing, and final cell parameter calculations were carried out using APEX4. These values are reported in the Results and Discussion. A numerical absorption correction was applied based on a Gaussian integration over a multifaceted crystal and followed by a semiempirical correction for adsorption applied using SADABS. The program SHELXT was used for the initial structure solution, and SHELXL was used for the refinement of the structure. Both of these programs were utilized within the OLEX2 software.

**RESULTS AND DISCUSSION**

**Pentapeptide Titration Curves.** Titration curves for the model pentapeptides were generated and used to assess the model PMF determined from BAR simulations as described in the methods. If the PMF is appropriately sampled and subtracted during constant-pH dynamics, then the predicted

<table>
<thead>
<tr>
<th>Identification code</th>
<th>CCDC no. 2292810</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical formula</td>
<td>$C_{11}H_{15}F_{3}N_{4}O_{5}$</td>
</tr>
<tr>
<td>Formula weight</td>
<td>340.27</td>
</tr>
<tr>
<td>Temperature/K</td>
<td>100</td>
</tr>
<tr>
<td>Crystal system</td>
<td>orthorhombic</td>
</tr>
<tr>
<td>Space group</td>
<td>$P2_1\overline{2}2_1$</td>
</tr>
<tr>
<td>$a$/Å</td>
<td>5.5240(2)</td>
</tr>
<tr>
<td>$b$/Å</td>
<td>13.2352(4)</td>
</tr>
<tr>
<td>$c$/Å</td>
<td>19.8946(5)</td>
</tr>
<tr>
<td>$\alpha$/deg</td>
<td>90</td>
</tr>
<tr>
<td>$\beta$/deg</td>
<td>90</td>
</tr>
<tr>
<td>$\gamma$/deg</td>
<td>90</td>
</tr>
<tr>
<td>Volume/Å³</td>
<td>1454.52(8)</td>
</tr>
<tr>
<td>$Z$</td>
<td>4</td>
</tr>
</tbody>
</table>

| $\rho_{\text{wG}}$/cm$^3$ | 1.554 |
| $\mu$/mm$^{-1}$         | 1.294 |
| F(000)                  | 704 |
| Crystal size/mm$^3$     | 0.138 $\times$ 0.075 $\times$ 0.047 |
| Radiation               | Cu Kα ($\lambda = 1.54178$) |
| $\Theta$ range for data collection/deg | 8.024 to 144.346 |
| Index ranges            | $-6 \leq h \leq 6, -16 \leq k \leq 16, -24 \leq l \leq 22$ |
| Reflections collected   | 17095 |
| Independent reflections | 2875 $[R_{\text{int}} = 0.0612, R_{\text{e}} = 0.0391]$ |
| Data/restraints/parameters | 2875/0/227 |
| Goodness of fit on F$^2$ | 1.064 |
| Final R indexes $[I \geq 2\sigma(I)]$ | $R_1 = 0.0391, wR_2 = 0.0996$ |
| Final R indexes $[\text{all data}]$ | $R_1 = 0.0428, wR_2 = 0.1021$ |
| Largest diff. peak/hole/e Å$^{-3}$ | 0.38/−0.29 |
| Flack parameter         | $-0.13(11)$ |

Figure 2. Thermal ellipsoids are represented at 50% probability. Carbon, hydrogen, nitrogen, oxygen, and fluorine atoms are represented by gray, white, light-blue, red, and light-green ellipsoids, respectively.
Table 3. The 11 Crystal Peptides Examined in This Study

<table>
<thead>
<tr>
<th>CCDC ID</th>
<th>Exp. Titration State</th>
<th>Space Group</th>
<th>Coformers</th>
<th>Asymmetric Unit Independent Predictions</th>
<th>Unit Cell Independent Predictions</th>
<th>Observed Fractions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAVZAQ (HID-MET)</td>
<td>P2₁</td>
<td>None</td>
<td>1</td>
<td>2</td>
<td>HId: &gt;0.99</td>
<td></td>
</tr>
<tr>
<td>RAVZEU (GLY-HID)</td>
<td>P2₁</td>
<td>Dihydrate</td>
<td>1</td>
<td>2</td>
<td>HIS: &gt;0.99</td>
<td></td>
</tr>
<tr>
<td>RAVZIY (LEU-HID)</td>
<td>P₂₁,2₂</td>
<td>Monohydrate</td>
<td>1</td>
<td>4</td>
<td>HIS: &gt;0.99</td>
<td></td>
</tr>
<tr>
<td>RAVZOE (HIS-ASP)</td>
<td>P2₁</td>
<td>Trihydrate</td>
<td>2</td>
<td>4</td>
<td>ASP: &gt;0.98</td>
<td></td>
</tr>
<tr>
<td>RAVZUK (HIE-GLH)</td>
<td>P2₁</td>
<td>None</td>
<td>2</td>
<td>4</td>
<td>HIE: &gt;0.99</td>
<td></td>
</tr>
<tr>
<td>RAWBAT (ALA-HIE)</td>
<td>P₂₁</td>
<td>Ethanol solvate hemihydrate</td>
<td>2</td>
<td>4</td>
<td>GLH: &gt;0.99</td>
<td></td>
</tr>
<tr>
<td>JUKMOR (HID-LEU)</td>
<td>P₂₁</td>
<td>None</td>
<td>1</td>
<td>2</td>
<td>HIS: &gt;0.99</td>
<td></td>
</tr>
<tr>
<td>TEKNAY (HIE-ALA)</td>
<td>P₂₁</td>
<td>Dihydrate</td>
<td>1</td>
<td>2</td>
<td>HIE: &gt;0.99</td>
<td></td>
</tr>
<tr>
<td>This Work (HIS-ALA)</td>
<td>P₂₁,2₂</td>
<td>Trithioacetic acid</td>
<td>1</td>
<td>4</td>
<td>HIS: &gt;0.99</td>
<td></td>
</tr>
<tr>
<td>CUFFUG (LYD)</td>
<td>P₂₁</td>
<td>None</td>
<td>2</td>
<td>4</td>
<td>LYD: 0.83</td>
<td></td>
</tr>
<tr>
<td>CURLOQ (CYD)</td>
<td>C2</td>
<td>Hexahydrate, tetrasodium</td>
<td>N.A.</td>
<td>8</td>
<td>CYD: &gt;0.99</td>
<td></td>
</tr>
</tbody>
</table>

Total Predictions | 14 | 40 |

The first column lists each crystal’s accession ID from the CCDC, followed by its space group in the second column. The third column lists coformer(s) present in the crystal along with the peptide of interest. In the fourth and fifth columns, the number of independent observations in the asymmetric unit and unit cell, respectively, are given. In the final column is the three-letter shorthand for the experimental titration state(s) and the observed titration state fractions from simulation. In the reported fractions, “Inter” represents the fraction of samples observed in an intermediate alchemical lambda state that is not considered to be consistent with a physical end-state. This value is listed only for the case of CUFFUG (LYD), where it was greater than for 1% of the samples. The reported space group for CCDC ID RAWBAT is P₂₁,2₂; however, it was expanded to P₂₁ for our simulations of the asymmetric unit. The asymmetric unit could not be appropriately simulated for CURLOQ due to the presence of atoms with fractional occupancy.
Table 4. Coordinate RMSDs and Distance Comparisons for the 11 Studied Peptides

<table>
<thead>
<tr>
<th>CCDC ID (Exp. Titration State)</th>
<th>Titration Site to Nearest Polar Contact</th>
<th>Experimental Distance (Å)</th>
<th>Mean Simulated Distance (Å)</th>
<th>RMSD (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAVZAQ (HID-MET) Histidine-Nδ-O-C terminus</td>
<td>2.72</td>
<td>2.96 ± 0.10</td>
<td>0.18 ± 0.03</td>
<td></td>
</tr>
<tr>
<td>RAVZEU (GLY-HID) Histidine-Nδ-O-C terminus</td>
<td>2.88</td>
<td>2.98 ± 0.09</td>
<td>0.16 ± 0.04</td>
<td></td>
</tr>
<tr>
<td>RAVZIY (LEU-HID) Histidine-Nδ-O-C terminus</td>
<td>2.73</td>
<td>2.90 ± 0.08</td>
<td>0.26 ± 0.05</td>
<td></td>
</tr>
<tr>
<td>RAVZOE (HIS-ASP) Histidine-Nε-O-Aspartate</td>
<td>2.63</td>
<td>2.85 ± 0.11</td>
<td>0.56 ± 0.10</td>
<td></td>
</tr>
<tr>
<td>RAVZUK (HIE-GLH) Histidine-Nε-O-C terminus</td>
<td>2.84</td>
<td>2.88 ± 0.07</td>
<td>0.19 ± 0.04</td>
<td></td>
</tr>
<tr>
<td>RAWBAT (ALA-HIE) Histidine-Nε-O-Water</td>
<td>2.87</td>
<td>2.97 ± 0.13</td>
<td>0.14 ± 0.03</td>
<td></td>
</tr>
<tr>
<td>JUKMOR (HID-LEU) Histidine-Nδ-O-C terminus</td>
<td>2.71</td>
<td>2.95 ± 0.09</td>
<td>0.26 ± 0.08</td>
<td></td>
</tr>
<tr>
<td>TEKNAY (HIE-ALA) Histidine-Nε-O-Water</td>
<td>2.80</td>
<td>2.86 ± 0.07</td>
<td>0.24 ± 0.08</td>
<td></td>
</tr>
<tr>
<td>This Work (HIS-ALA) Histidine-Nε-O-C terminus</td>
<td>2.66</td>
<td>2.76 ± 0.09</td>
<td>0.78 ± 0.17</td>
<td></td>
</tr>
<tr>
<td>CUFFUG (LYD) Lysine-Nζ-Nζ-Lysine</td>
<td>3.14</td>
<td>3.64 ± 0.48</td>
<td>0.60 ± 0.09</td>
<td></td>
</tr>
<tr>
<td>CURLOQ (CYD) Cysteine-S–Zn²⁺ ion</td>
<td>2.28</td>
<td>2.16 ± 0.04</td>
<td>0.38 ± 0.07</td>
<td></td>
</tr>
</tbody>
</table>

“The RMSDs were calculated using the progressive alignment of crystals (PAC) method, with a single peptide molecule used for the comparison. Coformers were ignored in the RMSD comparison. Interatomic distance comparisons were taken between the titrating heavy atom and its nearest polar contact, which may also be a titrating heavy atom. The standard deviations are based on the fluctuations observed during CpHMD simulations.

Figure 4. Representative interatomic distances for 4 of the 11 dipeptides encompassing each functional group titrated in the AMOEBA CpHMD simulations. (A) Interatomic distance between the nitrogens of lysine peptides in CUFFUG is shown. (B) CURLOQ system is shown with the interatomic distance between the sulfur of the cysteine side chain and the zinc ion. (C, D) Distance between the titrating ε-nitrogen of histidine and either the aspartate side chain of RAVZOE in the first case or the oxygen of the C-terminus of RAVZUK in the second case.

Figure 5. Alternative hydrogen bonding networks for histidyl alanine dihydrate tautomers (CCDC ID: TEKNAY) observed during the CpHMD simulations.
focused on histidine. The chemical structures for the 11 peptides can be seen in Figure 3. The cysteine crystal includes zinc, which is coordinated by two sulfur groups and two amino groups. This is particularly interesting as a model of zinc coordination in proteins, where zinc is often coordinated by cysteine.

As can be seen in Table 3, the AMOEBA constant-pH algorithm performs well in selecting the correct titration and tautomeric states while maintaining close agreement with the experimental structure, as seen in Table 4. Close attention was paid to the area around the titrating site, and no significant disturbances were found between polar contacts. Examples of this can be seen in Figure 4.

It is encouraging to see HID predictions that match experiment in each of the four cases, considering that in water, the HIE tautomeric state should be favored over HID roughly 70% of the time when histidine is deprotonated. The CpHMD AMOEBA algorithm matched the experiment in selecting the histidyl aspartate crystal as a salt and the histidyl glutamate as a cosubstrate with the correct tautomeric state. The prediction of deprotonated lysine and cysteine tautomeric states confirms that the model responds to net charge. Convergence was assessed from the separate asymmetric unit and P1 simulations. For all but TEKNAY, each system displayed a rapid convergence to the experimentally observed state. The asymmetric unit and P1 systems were nearly identical, and the results across the independently titrating sites in the P1 systems were also nearly identical. Each system converged within a few hundred picoseconds as demonstrated in the time series for the HIS-ALA TFA given in Figure S1 of the Supporting Information. This can also be observed from the populations given in Table 3.

The only compound for which the prediction did not strictly match the experiment was for the histidyl alanine dihydrate, TEKNAY. For this dipeptide, the CpHMD simulations sampled both the HID and HIE tautomeric states in significant populations. Over the course of the simulation, the histidine frequently sampled both deprotonated tautomeric states, HID and HIE, with a slight preference for HID. An analysis of the trajectory showed that each tautomeric state established a hydrogen bonding network with its surroundings that required minor conformational adjustments of the water molecules. As can be seen in Figure 5, the relative oxygen positions of the water molecules moved only a small amount to accommodate the reorientation of the hydrogen bond network. The distances between hydrogen bond donors and acceptors are also comparable. These observations suggest that both states are physically accessible and similar in energy. This is further supported by a free-energy assessment. Using the Rao–Blackwell estimator for free-energy differences as laid out by Ding et al., the free-energy difference between the two states from a 40 ns trajectory was calculated to be only 0.35 ± 0.03 kcal/mol.

**CONCLUSIONS**

We presented the first implementation, parametrization, and validation of a continuous constant-pH molecular dynamics algorithm using the multipolar polarizable AMOEBA force field under PME. Parameters were derived for the pentapeptides (Ace-AA-X-AA-Nme, where X is the residue of interest) for ASP, GLU, HIS, LYS, and CYS. These parameters were validated by constructing titration curves for each pentapeptide using the CpHMD algorithm accelerated with pH replica exchange. Further validation was performed on a set of 11 crystal peptides, 10 previously studied and 1 generated here. These crystal peptides provide a novel test suite for CpHMD methods using PME electrostatics because their atomic coordinates and titration states are generally well-defined. Each crystal was simulated as an asymmetric unit using FFX’s unique capability to handle symmetry operators as well as in the expanded unit cells (P1). The results were exactly replicated between the asymmetric unit and unit cell systems. The experimental titration states observed for each crystal peptide were reproduced for all except one crystal. Our simulations, supported by a free-energy difference estimate, indicated that disorder was present between the two tautomeric states of histidine for the histidyl alanine dihydrate crystal (CCDC ID: TEKNAY). Further investigation of this crystal system is ongoing, including recrystallization and refinement. The success of this algorithm in predicting titration states of peptides in a crystalline environment using the traditional parametrization of a peptide in water suggests that this model will be suitable for pK_a prediction in proteins under PME. Future work will incorporate CpHMD into AMOEBA protein simulations based on the generalized Kirkwood implicit solvent.

**ASSOCIATED CONTENT**

Data Availability Statement

Constant-pH molecular dynamics trajectories generated during this work can be regenerated using Force Field X (https://github.com/SchniedersLab/forcefieldx, https://ffx.biochem.uiowa.edu/).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c01180.

Functional form for each extended Hamiltonian equation and their parameters for the titrating functional groups; exact modifications to the AMOEBA protein force field parameters; and various figures providing additional context to the titration curves and convergence of the crystal simulations (PDF)

**AUTHOR INFORMATION**

Corresponding Author

Michael J. Schnieders — Department of Biomedical Engineering and Department of Biochemistry, University of Iowa, Iowa City, Iowa 52242, United States; orcid.org/0000-0003-1260-4592; Email: michael-schnieders@uiowa.edu

Authors

Andrew C. Thiel — Department of Biomedical Engineering, University of Iowa, Iowa City, Iowa 52242, United States
Matthew J. Speranza — Department of Biomedical Engineering, University of Iowa, Iowa City, Iowa 52242, United States
Sanika Jadhav — Department of Pharmaceutical Sciences and Experimental Therapeutics, University of Iowa, Iowa City, Iowa 52242, United States
Lewis L. Stevens — Department of Pharmaceutical Sciences and Experimental Therapeutics, University of Iowa, Iowa City, Iowa 52242, United States; orcid.org/0000-0003-2936-0926
Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c01180

Funding
A.C.T. was supported by an NSF (National Science Foundation) Graduate Research Fellowship under grant no. 000390183. J.W.P. and P.R. were supported by NIH grants R01GM114237 and R01GM106137. J.S. was supported by NIH grant R35GM148261. P.R. thanks support from the Welch Foundation (F-2120). M.J.S. was supported by NIH grant R35GM148261 and NSF grant CHE-1751688.

Notes
The authors declare the following competing financial interest(s): J. Ponder and P. Ren are cofounders of Qubit Pharmaceuticals.

REFERENCES

(32) Huang, Y.; Chen, W.; Wallace, J. A.; Shen, J. All-atom continuous constant pH molecular dynamics with particle mesh


(57) APEX, B. APEX4, SAINT, and SADABS; Bruker AXS Inc.: Madison, WI, 2021.


