JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC

Article

Accurately Predicting Protein pK_a Values Using Nonequilibrium Alchemy

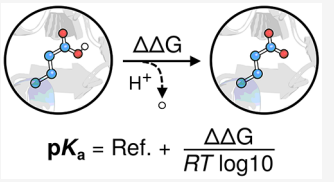
Carter J. Wilson, Mikko Karttunen, Bert L. de Groot, and Vytautas Gapsys*

Cite This: J. Chem. Theory Comput. 2023, 19, 7833−7845

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The stability, solubility, and function of a protein depend on both its net charge and the protonation states of its individual residues. pK_a is a measure of the tendency for a given residue to (de)protonate at a specific pH. Although pK_a values can be resolved experimentally, theory and computation provide a compelling alternative. To this end, we assess the applicability of a nonequilibrium (NEQ) alchemical free energy method to the problem of pK_a prediction. On a data set of 144 residues that span 13 proteins, we report an average unsigned error of 0.77 ± 0.09, 0.69 ± 0.09, and 0.52 ± 0.04 pK for aspartate, glutamate, and lysine, respectively. This is comparable to current state-of-the-art predictors and the accuracy recently reached using free energy perturbation methods (e.g., FEP+). Moreover, we demonstrate that our open-source, `pmx`-based approach can accurately resolve the pK_a values of coupled residues and observe a substantial performance disparity associated with the lysine partial charges in Amber14SB/Amber99SB*-ILDN, for which an underused fix already exists.



INTRODUCTION

Amino acids with ionizable side chains make up approximately 30% of the residues found in proteins[1,2] and play a key role in maintaining protein stability,[3−6] modulating solubility,[7,8] mediating protein−protein interactions,[9,10] and facilitating cell signaling.[11,12] These amino acids, namely, aspartate, glutamate, arginine, lysine, cysteine, tyrosine, and histidine, are functionally dependent on their protonation states, which vary depending on their local environments. The measure of this dependence is known as the pK_a, which relates the pH of the solution to the protonation state of a residue via the Henderson−Hasselbalch equation, i.e., pK_a = pH + log[HA]/[A$^-$]. Given its degree of solvent exposure, Coulombic interactions, and hydrogen bonding, the pK_a of an amino acid residue may be raised or lowered relative to its reference pK_a°—determined using a capped peptide (e.g., ACE-AXA-NH$_2$) in solution—resulting in a lower or higher likelihood of protonation at a given pH. For acidic groups, the pK_a values tend to be elevated relative to their reference,[13−15] while for basic groups, the pK_a values tend to be lowered relative to their reference.[16,17] These shifts away from the reference value can reach up to ±5 pK units, and in many proteins, key ionizable residues are situated in such a way that a perturbation of their pK_a allows them to perform unique and specific functions.[18−22] The existence of such functional motifs relies on the alterable stability of the covalent bond between hydrogen and its heavy atom (e.g., O−H and N−H). The tendency of a side chain containing these groups to (de)protonate in a given microenvironment is quantified by the pK_a.

The relationship of protein−ligand binding to pK_a is of particular interest.[23−25] Here, the pK_as of both the ligand and the binding site residues as well as the pH- and binding-induced conformational changes of the protein are all intimately related. Resolving the precise states of the ionizable residues, as well as the local conformations of the apo and holo protein, are active fields of study that involve both experimental[26−28] and computational approaches.[29−32]

The conventional and often most precise method to determine the pK_a of an ionizable side chain is to measure the pH dependence of the main or side-chain chemical shifts using multidimensional nuclear magnetic resonance (NMR) spectroscopy.[33−35] The dependence of the chemical shift on pH is then fit to the Henderson−Hasselbalch equation, and the pK_a is resolved from the point of inflection. NMR can estimate the pK_a with an accuracy of 0.1−0.2 pK unit;[36] however, this strongly depends on the nuclei considered (i.e., ^{13}C vs ^{15}N) and the fit to the Henderson−Hasselbalch curve, which can be difficult due to conformational changes,[37] titration coupling,[38] or if the chemical shift simply reports a different titration event.[36,39] Even with the above caveats, NMR remains the experimental method of choice to resolve pK_a values in proteins and is, in general, very reliable. There are alternative approaches for measuring pK_a values including fluorometry, kinetic assays, and isothermal titration calorimetry.[40−42] However, they have their own challenges and generally obtain pK_a values with higher uncertainty compared to the NMR-based approach.

Received: June 29, 2023
Published: October 11, 2023

Theoretical methods are a compelling alternative to experiments. Many of these are motivated by a free energy formalism based on the thermodynamic cycle shown in Figure 1. Here, we consider a residue of interest (A) in both protein

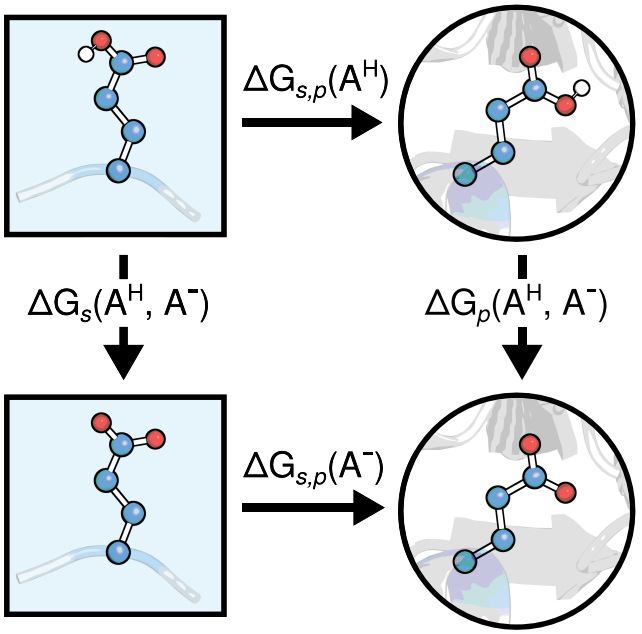


Figure 1. Thermodynamic cycle to compute the free energy difference between protonating a residue in a capped peptide in solution and the same residue in a protein. This $\Delta\Delta G$ can be related to pK_a(protein) given the reference pK_a° via eq 1.

(Figure 1, right) and reference peptide in solution (Figure 1, left). We assume that the reference pK_a° is known, then pK_a(protein) is given by

$$pK_a(\text{protein}) = pK_a^\circ + \frac{\Delta G_p(A^H, A^-) - \Delta G_s(A^H, A^-)}{RT\log(10)} = pK_a^\circ + \frac{\Delta\Delta G_{p,s}(A^H, A^-)}{RT\log(10)} \quad (1)$$

Note that $\Delta\Delta G_{p,s}(A^H, A^-)$ implicitly contains two terms. The first ($\Delta\Delta G^{env}$) represents the free energy of dissociating a proton within a protein relative to the reference state (e.g., capped peptide), where the protein residues are fixed to some state such that the value is pH independent; the second (ΔG^{titr}(pH)) accounts for the contribution from the titratable residues in the protein as they (de)protonate with pH. A consideration of the first component yields a $pK_{int} = pK_a^\circ + \frac{\Delta\Delta G^{env}}{RT\log(10)}$, which can be further used to calculate the true pK_a

$$pK_a(\text{protein}) = pK_{int} + \frac{\Delta G^{titr}(\text{pH})}{RT\log(10)} \quad (2)$$

In most cases, it is safe to assume that the mutual dependence (or coupling) of a residue A and its protein microenvironment is small, and by simply assigning residues to the charge state most likely for a corresponding model compound in solution (e.g., capped peptide) at pH ≈ 7.4, we can assume pK_a(protein) ≈ pK_{int}. However, there are cases where this assumption will fail, and a consideration of ΔG^{titr}(pH), at least for nearby titratable residues, is necessary to correctly resolve pK_a(protein). To that end, we have elsewhere introduced a thermodynamic-cycle-based formalism to account for this additional titration contribution and therein discuss the role of microscopic pK_a values in the context of coupled residues.[43]

Whether or not the pH dependence on the pK_a is taken into account, the fundamental aim of most theoretical methods is to resolve the free energy difference in eq 1 and thus estimate the pK_a. This can be done within a macroscopic or microscopic framework; we briefly describe both.

Macroscopic frameworks model the entire system, protein, and solvent as either a regularly shaped or an irregularly shaped object situated within a dielectric medium. From this, the energy terms can be resolved using the Poisson−Boltzmann equation (PBE). For a regularly shaped protein (e.g., idealized sphere), the PBE can be solved analytically;[44,45] however, for a more realistic, irregularly shaped protein, the PBE must be solved numerically. The numerical Poisson−Boltzmann (PB) approach for computing pK_a values was pioneered by Bashford and Karplus[46] and has since been continually refined.[47] Changes in both the underlying algorithmic and numerical formalism (e.g., parameter selection, linearized PBE,[48] etc.) and the structural descriptions of the system (e.g., partial charge changes,[49] side-chain rotamers,[50] etc.) have aimed to increase accuracy and applicability.

A microscopic framework based on atomistic simulations,[51] unlike a macroscopic one, in theory, does not require the definition of empirical parameters (e.g., charge density) or physical quantities (e.g., permittivity). The principal drawback is the computational cost that can be overcome by modification of the underlying model representation or implementation (e.g., the reintroduction of pseudoparameters) or by improvements in computing power. Molecular dynamics (MD) simulations offer an attractive solution for sampling biomolecular ensembles spanning meaningfully long time scales with fully atomistic representations of both protein and solvent. These simulations and the resultant ensembles might be used as an input for a PB-based approach,[52−55] or can be performed in conjunction with a free energy method (e.g., thermodynamic integration,[56,57] free energy perturbation,[58] etc.), allowing for a direct resolution of the $\Delta\Delta G$ between protonation states. An alternative MD-based approach is constant pH molecular dynamics (CpHMD) simulations. Here, Monte Carlo sampling[59−62] (discrete CpHMD) or λ-dynamics[63−66] (continuous CpHMD) is used to explicitly sample protonation events. This allows for an explicit consideration of the proton concentration, where the protonation states of titratable residues are not restrained but are allowed to dynamically follow the free energy gradient.

Empirical (EM) approaches stand in contrast to those described above, which are primarily based on a rigorous free energy formalism. Empirical methods tend to rely on sets of approximate functional forms (e.g., hydrogen bonds) with knowledge-based parameters that are optimized based on large training sets of measured pK_a values. Such approaches have generated predictors with impressive accuracy at low computational cost,[67,68] which have been further enhanced with the advent of machine learning.[69,70]

It can be said that for all of the methods mentioned above, the objective is to provide predictive accuracy within the same range as that reached by experiment (i.e., <0.2 pK units). A perfect method ought to be system independent and hence not require fitting to experimental data. It should be able to

https://doi.org/10.1021/acs.jctc.3c00721
J. Chem. Theory Comput. 2023, 19, 7833−7845

robustly predict the free energies of protonation in the core of a protein and in the solvent-exposed regions, which requires that solute−solvent interactions be accurately represented. Moreover, the ability to change environmental conditions (e.g., temperature and ionic concentration) is another necessary requirement.

Alchemical free energy calculations based on molecular dynamics (MD) simulations have the potential to fulfill these requirements. Previous work has demonstrated that non-equilibrium (NEQ) free energy methods are able to accurately estimate the effects of mutations on protein stability,[71] as well as relative[72] and absolute protein−ligand binding affinities.[73] However, the ability to seamlessly and consistently extend these free energy frameworks to pH-dependent contexts, where invariably differences in the residue protonation states will measurably shift the computed free energies, and where assignment of the protonation states requires knowledge of the pK_a values, first requires a successful demonstration that plain pK_a values can be resolved using NEQ.

To this end, we use `pmx`-based NEQ free energy calculations to compute the $\Delta\Delta G$ and corresponding pK_a values (as described in eq 1) for 144 residues across 13 different proteins in two contemporary force fields. The calculated free energy differences were combined into a consensus estimate. We also consider six popular and well-validated alternative computational methods as a comparison. Additionally, we compare our results to pK_a values computed using FEP+[58] (Schrödinger Inc.) and observe no statistically significant difference between the accuracy achieved with both methods. We also report substantial performance disparities for lysine residues in Amber14SB,[74] which are caused by the partial charge assignment of the backbone and for which corrections already exist.[75] Furthermore, we demonstrate the ability of our `pmx`-based approach to accurately resolve the pH-dependent pK_a values of coupled residues, expanding the potential use for probing amino acids involved in unique redox or catalysis reactions. The average unsigned error (AUE) of the `pmx`−computed pK_a values across the residue classes considered was 0.68 ± 0.05 pK. The open-source `pmx` tool[76] is freely available at https://github.com/deGrootLab/pmx.

METHODOLOGY

Data Sets. The structures for the pK_a calculations were taken from the PDB database. Identifiers (and the corresponding experimental pK_a data) are as follows: 1BPI[77] (data[78,79]), 1BNR[80] (data[81]), 1BEO[82] (data[83]), 6QFS[84] (data[84]), 3BDC[85] (data[85]), 1CLB[86] (data[87,88]), 1RGG[89] (data[90]), 2LZT[91] (data[36]), 4TRX[92] (data[93]), 2RN2[94] (data[95]), 1OMU[96] (data[97]), 1NZP[98] (data[99]), and 1LKJ[100] (data[18]) (see the Supporting Information (SI) for details about 1LKJ). The list of proteins, their residues, and the corresponding experimental pK_a values are provided in Table S1.

PDB structure IDs for thermostability calculations and the corresponding experimental $\Delta\Delta G$ data references are as follows: 1EY0[101] (data[17,102]), 2LZM[103] (data[104−110]), and 2RN2[94] (data[111]). The list of proteins, their residues, and the corresponding experimental $\Delta\Delta G$ values are provided in Table S2.

We make reference to four main pK_a data sets:

- full: 13 proteins and 144 residues: 57 aspartate, 48 glutamate, and 39 lysine residues (main data set used for method comparison; all other data sets are subsets)
- FEP+: contains the 65 residues that overlap with a recent FEP+ publication[58] (used to compare NEQ and FEP+ approaches)
- lysine: contains 13 lysine residues from hen egg-white lysozyme (HEWL) and calbindin 9k (used to assess the source of a lysine performance discrepancy)
- reduced: contains 15 aspartate and 14 glutamate residues from SNase + ΔPHS and HEWL (used to assess Amber99SB-*disp* performance)

Nonequilibrium Alchemy. `pmx`[76] was used for the system setup, hybrid structure and topology generation, and analysis. Initial structures were taken from the PDB database (see the Methodology section).

A double system in a single box setup was used; here, both the protein and peptide (e.g., ACE-AXA-NH_2) are situated at a distance of 3 nm in the same box, which ensures charge neutrality during the alchemical transition.[112] To prevent consequential protein−peptide interactions, a single Cα in each molecule was positionally restrained. Given the thermodynamic cycle used (Figure 1), the free energy cost associated with this restraint cancels between the two vertical branches. We used the CHARMM36m[113] (with CHARMM-modified TIP3P[114]) and Amber14sb[74] (with TIP3P[115]) force fields.

For all systems, an initial minimization was performed by using the steepest descent algorithm. A constant temperature corresponding to the reference experimental setup was maintained implicitly using the leapfrog stochastic dynamics integrator[116] with an inverse friction constant of $\gamma = 0.5$ ps^{-1}. Pressure was maintained at 1 bar using the Parrinello−Rahman barostat[117] with a coupling time constant of 5 ps. The simulation time step was set to 2 fs. Long-range electrostatic interactions were calculated using the particle-mesh Ewald method[118] with a real-space cutoff of 1.2 nm and a Fourier spacing of 0.12 nm. Lennard-Jones interactions were force-switched off between 1.0 and 1.2 nm. Bonds to hydrogen atoms were constrained using the Parallel LINear Constraint Solver.[119]

To improve sampling, systems were run for 25 ns in four independent replicas; in each case, the first 5 ns were discarded as equilibration. From the remaining 20 ns, 200 non-equilibrium transitions of 200 ps were generated and work values from the forward and backward transitions were collected using thermodynamic integration. These values were then used to estimate the corresponding free energy with Bennett's acceptance ratio[120] as a maximum likelihood estimator relying on the Crooks fluctuation theorem.[121] Bootstrapping was used to estimate the uncertainties of the free energy estimates,[112,122] and these were propagated when calculating the $\Delta\Delta G$ values. By varying the length of equilibrium and transition simulations as well as the number of transitions, we ensured that this simulation protocol yields converged free energy estimates (Figure S1). Equation 1 was used to convert between $\Delta\Delta G$ and pK_a(protein) values using the corresponding references (i.e., aspartate: 3.94 ± 0.03, glutamate: 4.25 ± 0.05, and lysine: 10.4 ± 0.08).[123,124]

Conventional Predictors. In addition to the MD-based pK_a estimation, we also considered an empirical (EM) method PropKa[67,68] (v3.4); four Poisson−Boltzmann (PB) methods:

https://doi.org/10.1021/acs.jctc.3c00721
J. Chem. Theory Comput. 2023, 19, 7833−7845

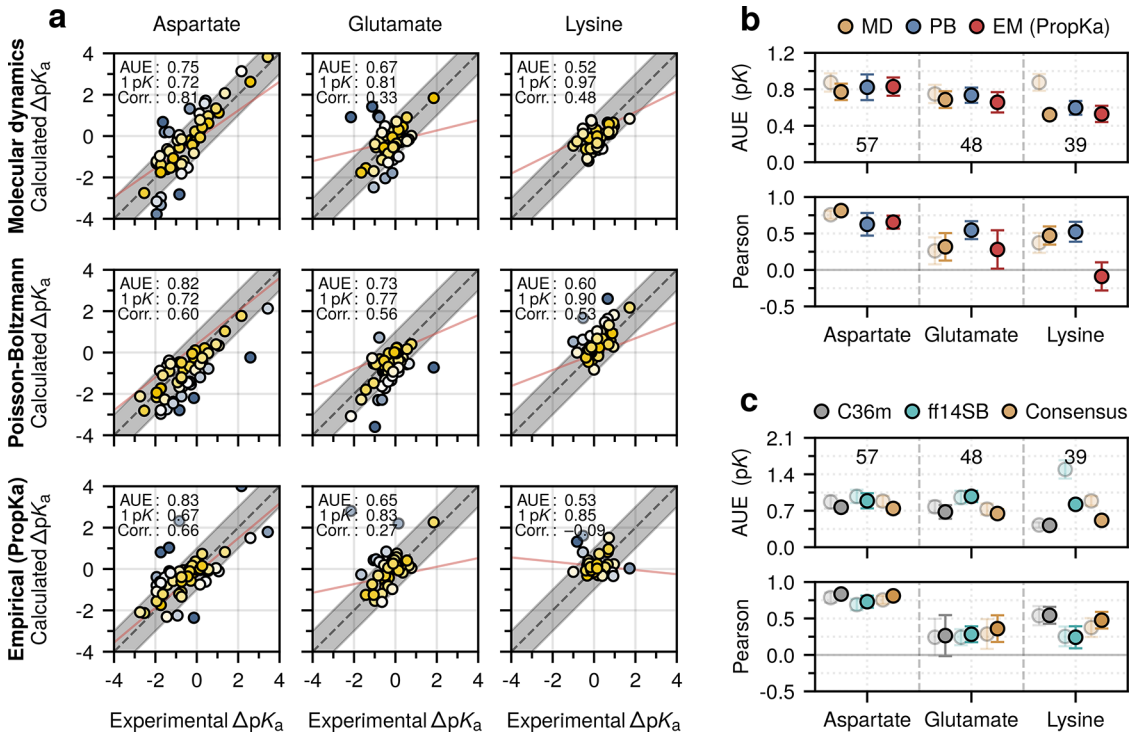


Figure 2. Full data set residue-wise performance. (a) Correlation between the calculated and experimental pK_a values. MD values are adjusted for residue coupling and lysine parametrization. Marker color indicates deviation from experiment. Regression lines are indicated in red. The proportion of residue 1 pK units from experiment is indicated. (b) Average unsigned errors (AUEs) and Pearson correlation coefficients computed for the various methods: molecular dynamics (MD), Poisson−Boltzmann (PB), and the empirical PropKa approach (EM). (c) AUEs and Pearson correlation coefficients were computed for the two force fields: CHARMM36m and Amber14SB, and their consensus. Transparent markers indicate the unadjusted estimates. Numerical values indicate the number of residues considered. When available, bootstrapped standard errors are depicted.

DelPhiPKa[125,126] (v2.3), H++[127] (v4.0), MCCE[50,128] (v2.8), and PypKa[129] (v2.9.4); and evaluated a machine-learning-based predictor pK_a-ANI[70] (v.0.1.0).

PropKa is an empirical predictor, where the ΔG contributions are captured by Coulombic, desolvation, and intrinsic electrostatic (e.g., hydrogen bonding) energy equations. Default settings were used when performing the calculations.

DelPhiPKa, as with all PB methods considered here, calculates the electrostatic potential by numerically solving the PBE using a finite difference method. Based on DelPhi software, this method uses a smooth Gaussian function to capture the heterogeneous dielectrics of the solute and solvent. Default settings were used except for the salt concentration, which was set according to the experimental setup (Table S1).

H++ relies on the single-conformer version of MEAD[130] and assigns charges and parameters based on Amber99SB. Default settings were used except for the default pH, which was set to 7.4, and the salt concentration, which was set according to the experimental setup (Table S1).

MCCE, based on DelPhi, uses Monte Carlo simulations to capture dynamic side-chain conformational changes. Default settings were used except for the salt concentration, which was set according to the experimental setup (Table S1).

PypKa uses Monte Carlo calculations to probe the proton tautomers and employs DelPhi to solve the PBE. Default settings were used, except for the salt concentration, which was set according to the experimental setup (Table S1).

pK_a-ANI can also be considered an empirical predictor. This predictor utilizes deep representation learning[131] that combines an atomic environment vector and the neural network potential ANI-2x.[132] Default settings were used when performing the calculations, including a gas-phase minimization of the initial PDB structures in GROMACS using the Amber14SB force field.

RESULTS

Overall Performance. Double free energy differences ($\Delta\Delta G$) were calculated for all 144 residues (48 aspartates, 57 glutamates, and 39 lysines), allowing us to robustly evaluate performance on a large data set. For the MD-based and PB-based approaches, a consensus estimate was used to make comparison easier. The EM-based approach corresponds to PropKa calculations, while the ML-based pK_a-ANI method is discussed in a separate section.

With respect to the MD approach, we observed two important sources of prediction inaccuracy: residue coupling and lysine parametrization. Adjusting the pK_a calculation framework to account for these led to an adjusted estimate that we compare to the unadjusted one. This is extensively discussed in the Determinants of Accuracy: Lysine Parametrization and Determinants of Accuracy: Protonation Neighborhood and Residue Coupling sections.

Figure 2 summarizes the main findings: in absolute terms, MD-based nonequilibrium free energy calculations perform comparably to conventional in silico predictors, with an overall adjusted predictive AUE of 0.68 ± 0.05 pK taken as an average over each residue class (compared to 0.74 ± 0.07 and 0.70 ± 0.06 pK for the consensus of the Poisson−Boltzmann (PB) methods and empirical (EM) PropKa method, respectively)

https://doi.org/10.1021/acs.jctc.3c00721
J. Chem. Theory Comput. 2023, 19, 7833−7845

(Figure 2a,b). Regarding the individual residue classes computed using the MD approach, for the adjusted estimate, AUEs were 0.77 ± 0.09 pK (aspartate), 0.69 ± 0.09 pK (glutamate), and 0.52 ± 0.04 pK (lysine) (Figure 2a,b).

The unadjusted force-field differences revealed that CHARMM36m performed as well or better for each residue class compared to Amber14SB (Figure 2c). The most notable differences were evident for lysine, where Amber14SB significantly underperformed compared to CHARMM36m (AUE: 0.42 ± 0.05 vs 1.48 ± 0.18 pK).

The Pearson correlation coefficients revealed a similar trend; for aspartate and lysine, the adjusted MD-based estimate gave values of 0.81 ± 0.04 and 0.48 ± 0.12, respectively, performing as well or better than the alternative approaches (PB: 0.61 ± 0.16 and 0.52 ± 0.13; EM (PropKa): 0.67 ± 0.08 and −0.09 ± 0.19). For glutamate, weaker correlations with the MD-based approach (0.33 ± 0.19) were evident. Regardless of the method, the highest correlations were for aspartate, where the experimental pK_a values had the largest dynamic range, while the weakest correlations were for lysine, where the dynamic range of the experimental values was narrower (Figure 3).

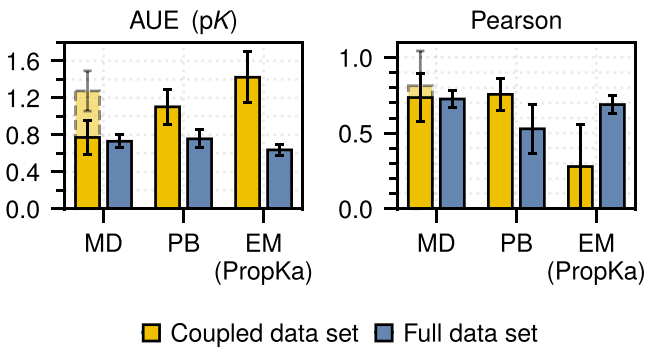


Figure 3. Performance of methods on coupled residues. Average unsigned errors (AUEs) and Pearson correlation coefficients were computed for both the coupled data set (i.e., 18 aspartates and glutamates) and the full data set aspartate and glutamate residues, with the coupled set discarded. Dashed lines indicate the performance of the MD-based approach before coupling was accounted for (see text). Bootstrapped standard errors are depicted.

We did not observe a strong dependence of the prediction accuracy on the protein system. Rather, the systems for which higher accuracy was observed (Figure S2) contained a higher proportion of probed lysine residues (e.g., 1NZP and 1LKJ), again illustrating disparate pK_a prediction accuracy for different residue types. In general, residues with larger ΔpK_a values (Figure S3) and lower solvent exposure (Figure S4) tended to be predicted worse. We note that these two variables are related: probed residues with smaller ΔpK_as were also found to be more solvated (Figure S5).

Determinants of Accuracy: Lysine Parametrization. As discussed above, Amber14SB provided markedly poorer estimates of the $\Delta\Delta G$ compared with CHARMM36m for most of the lysine residues considered, significantly underestimating the pK_a values (Figure 4a,d). We conceived of two potential sources of error: (1) environmental and (2) residue parametrization. Given the discussions in the literature pertaining to ion overbinding[133−135] and the role of a solvent model on protein solvation,[136] we began by assessing the role of environmental conditions. Specifically, we probed K^+ (rather than Na^+) counterions, NBFIX parameters,[134] Åqvist[137] (rather than Joung/Cheatham[138]) ion parameters, and TIP4P-D water[139] (rather than TIP3P). Using these variants, the pK_a values of lysines from a 13 residue data set (i.e., hen egg-white lysozyme (HEWL) and calbindin 9k) were computed. No significant improvement in the estimates was observed (Figure 4a,c).

To consider the role of parametrization, simulations were performed with three different versions of Amber, namely, Amber99SB*-ILDN,[140−142] Amber03*,[141,143] and Amber99SB-*disp*.[144] On the same lysine data set, a dramatic improvement was observed with Amber99SB-*disp* (Figure 4a,c). Given that differences in the dihedral parametrization between Amber99SB*-ILDN and Amber14SB appeared to confer almost no performance improvement, this narrowed the likely cause of the difference to the nonbonded interactions. Regarding the Lennard-Jones terms, Amber99SB-*disp* alters the parameters of aspartate, glutamate, and arginine, leaving open the possibility of more accurate interactions between lysine and other charged residues in the protein as the source of this discrepancy. However, more notable was the inclusion of the Best et al. lysine partial charges (i.e., Amber99SB*-ILDN-Q[75]) with Amber99SB-*disp*. Although both Amber14SB and Amber99SB*-ILDN have the same partial charge assignment, Amber99SB-*disp* uses altered backbone charges for aspartate, glutamate, lysine, arginine, and doubly protonated histidine (Figure 4b). These were originally developed in the Amber99SB*-ILDN-Q force field to correct for aberrant helical propensities and create consistency among the amino acids. In both Amber99SB*-ILDN and Amber14SB, with the exception of proline, all but these five charged residues have the same assigned backbone partial charge set for C, O, N, and HN. By using the updated parameters by Best et al., both protonated (LYS) and deprotonated (LYN) lysine in Amber99SB*-ILDN-Q and Amber99SB-*disp* have the same charge assignment for C, O, N, and HN.

Such a backbone partial charge assignment is akin to that in the CHARMM36m force field, which has the same backbone partial charge sets (including the Cα and Hα atoms) for all residues except proline and glycine.

We constructed a hybrid Amber14SB-K force field with the altered lysine partial charges but only for the probed residue. We found that this force field performed markedly better on the lysine data set, cutting the average unsigned error by almost half, from 1.48 ± 0.18 to 0.81 ± 0.08 pK (Figure 4a,c). The improvement was most pronounced for lysine residues in the helical regions (Figure S6). This result, in addition to that from Best et al.,[75] suggested that the default partial charges of lysine were erroneous. To further assess the effect of partial charges, we computed the thermostability of 15 lysine mutations using CHARMM36m, Amber14SB, and Amber14SB-K. We again observed a marked improvement in the AUE using the altered lysine partial charges, which shifted the value from 10.42 to 5.54 kJ/mol (Figure 4e).

While Amber99SB-*disp* exhibited the highest accuracy on the lysine data set (Figure 4c), suggesting its general use for pK_a prediction, this behavior did not hold for aspartate and glutamate. On a reduced data set (i.e., SNase + ΔPHS and HEWL), Amber99SB-*disp* exhibited below-average accuracy (Figure S7).

Determinants of Accuracy: Protonation Neighborhood and Residue Coupling. Overall, alchemical free energy calculations and conventional pK_a predictors provide comparable accuracy. However, unlike many alternative

https://doi.org/10.1021/acs.jctc.3c00721
J. Chem. Theory Comput. 2023, 19, 7833−7845

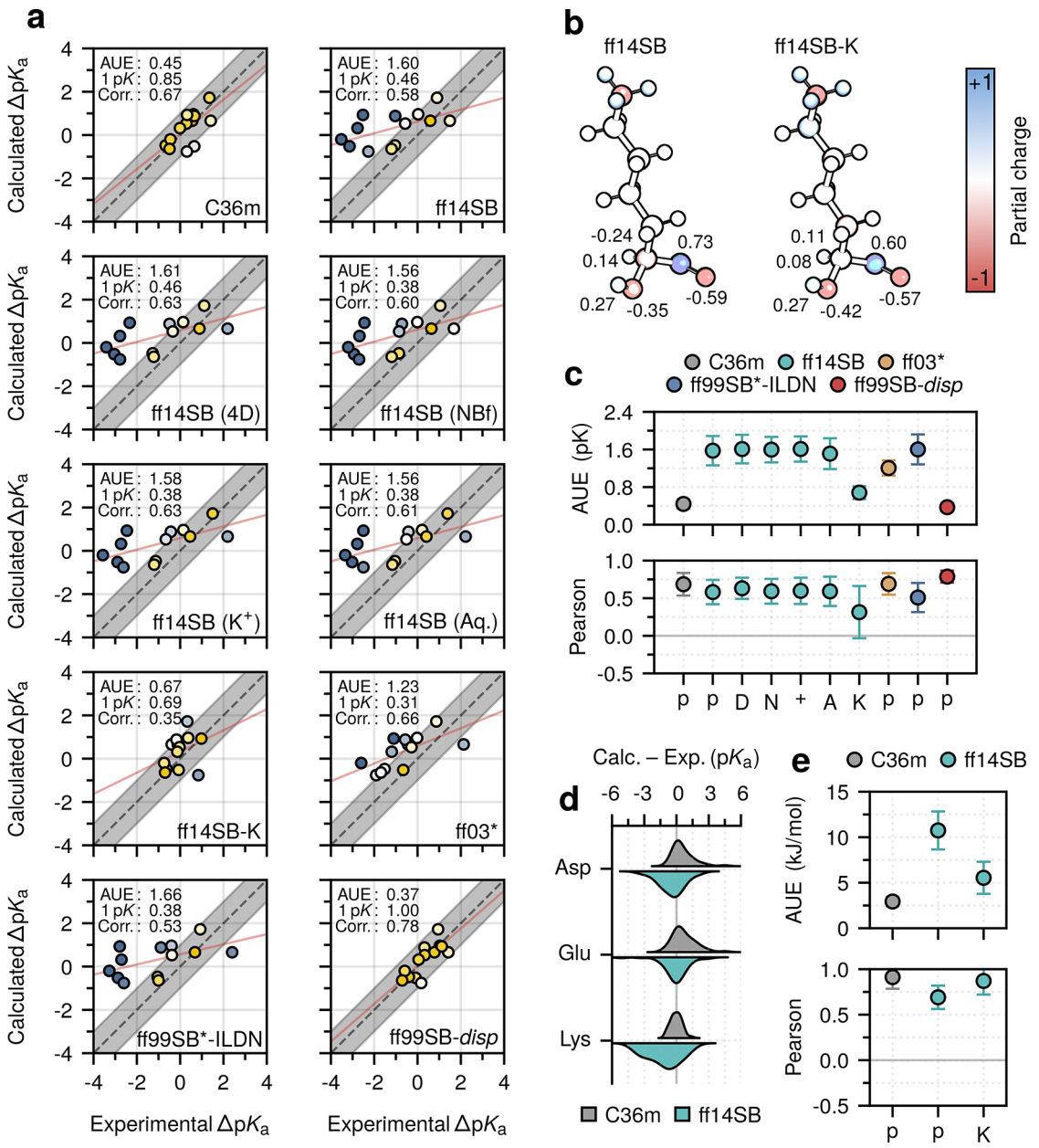


Figure 4. Calculating lysine pK_a values with different force fields. (a) Correlation between the calculated and experimental pK_a values. Marker color indicates deviation from experiment. Regression lines are indicated in red. The proportion of residues 1 pK unit from experiment is indicated. (b) Partial charge assignment differences between Amber14SB and Amber14SB-K. Numeric values correspond to backbone atoms. (c) Average unsigned errors (AUEs) and Pearson correlation coefficients computed for the various force-field combinations: five variants of Amber14SB (with TIP4P-D (D), with NBFIX (N), with K^+ counterions (+), with Åqvist ions (A), or with Best et al. charges assigned to the probed lysine (K)), as well as "plain" (p) CHARMM36m, Amber14SB, Amber03*, Amber99SB*-ILDN, and Amber99SB-*disp*. (d) Distribution of differences between the unadjusted MD-based and experimental pK_a values. (e) AUEs and Pearson correlations computed on a lysine thermostability data set. When available, bootstrapped standard errors are depicted.

approaches, the alchemical method described here allows for the resolution of conditional pK_a values. The consideration of such values may not only improve the estimates but also allow one to determine the pH-dependent pK_a of a residue. Recently, we derived a formalism to conveniently combine double free energy differences from alchemical calculations in order to account for coupling between residues when predicting the pK_a.[43]

In this work, we selected 18 residues, including several acidic dyads across the data set, for which the deviation from experiment was >1 pK. We further calculated the pK_a values of these residues by taking into account possible couplings with the protonatable residues in their neighborhood. For residues neighboring a histidine, standard pK_a calculations were performed in the presence of doubly protonated histidine, i.e., we assume this to be the protonation state at the pH where aspartate and glutamate titrate. For pairs of nearby (i.e., <0.5 nm) acidic residues, we applied the aforementioned thermodynamic formalism, while for apparent triads, an assessment of the most probable deprotonation event was first determined, followed by an application of the formalism on the remaining dyad. Explicitly accounting for residue

https://doi.org/10.1021/acs.jctc.3c00721
J. Chem. Theory Comput. 2023, 19, 7833−7845

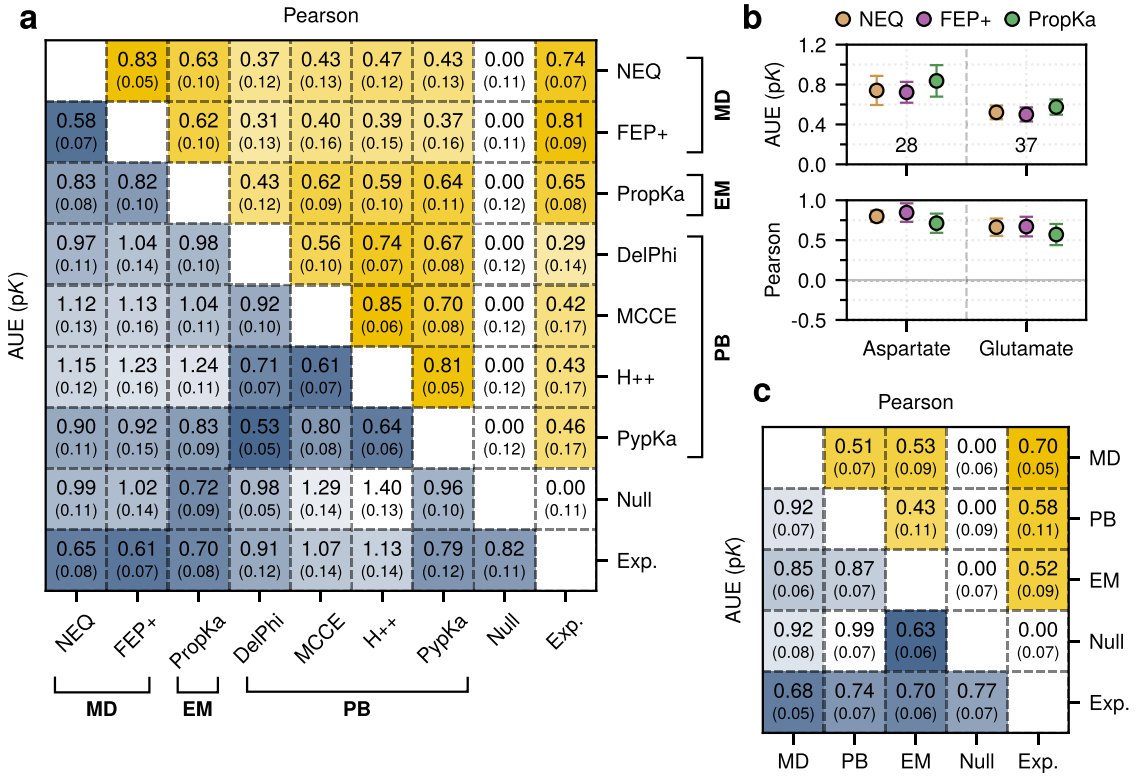


Figure 5. Comparison of the ΔpK_a predictions by each method. (a) Pearson correlations (upper right triangle) and AUEs (lower left triangle) between ΔpK_a estimates were calculated for each method over the FEP+ data set. Comparison with experiment means that the bottom row and rightmost column correspond to the overall performance. DelPhiPKa is abbreviated DelPhi. (b) Individual residue-wise error plot of the NEQ, FEP+, and PropKa methods on the FEP+ data set. Numerical values (i.e., 28 and 37) indicate the number of residues considered. (c) Pearson correlations and AUEs for the three ΔpK_a consensus estimates were calculated over the full data set; note that EM corresponds to PropKa. Comparison with experiment means that the bottom row and rightmost column correspond to overall performance. Bootstrapped standard errors are indicated.

coupling reduced the AUE from 1.28 to 0.76 pK of the residues considered, bringing the accuracy close to the AUE observed over the full data set (Figure 3). For all of the methods considered, this coupled residue subset had higher errors than those observed on the remaining aspartate and glutamate residues (i.e., full data set minus coupled subset).

We note that this analysis was retrospective, where we have a priori access to the correct pK_a values, i.e., we could preselect which residues to subject to these more involved calculations involving inter-residue couplings. However, in principle, such calculations can be applied to any residues with nearby protonatable neighbors. Our formalism[43] ensures that if alchemical calculations suggest no coupling, the final pK_a estimate will remain similar to that of a standard calculation without coupling considerations.

Method Comparison. Recently, FEP+ was used to compute the pK_a values of 79 aspartate and glutamate residues.[58] We observed comparable performance on the overlapping 65 residue data set (referred to as the FEP+ data set); the average unsigned error was 0.65 ± 0.08 for NEQ and 0.61 ± 0.07 for FEP+ (Figure 5a), and the Pearson correlation coefficients were 0.74 ± 0.06 and 0.80 ± 0.09, respectively. These represented the two strongest performing methods on the FEP+ data set. We also assessed the degree of correlation between the ΔpK_a estimates for both methods; here, the Pearson correlation coefficient was 0.83 ± 0.05, suggesting a strong relationship (Figure 5a). This was the second strongest correlation between any two methods on the FEP+ data set. Regarding residues, glutamate pK_a values were predicted with a higher accuracy than aspartate (Figure 5b).

We also considered our NEQ approach in relation to individual computational methodologies (rather than a consensus), including the popular PropKa software. Given the computational efficiency of this empirical method, it presents a compelling approach for large-scale pK_a calculations. We found that NEQ and FEP+ could outperform PropKa on the FEP+ data set (Figure 5a,b); however, PropKa still showed strong performance on the full data set (Figures S8 and S9). For the full data set, while the AUE values for PropKa predictions were small, the correlations also tended to be weaker. This was particularly evident for lysine, where the Pearson correlation coefficient was near zero. For the precise discrimination of individual residues and an absolute ordering of pK_a values, an MD-based free energy approach may be warranted.

As with FEP+, we evaluated the degree of correlation and deviation between the ΔpK_a values computed using various methods. The strongest correlations were observed within method classes (e.g., DelPhiPKa/MCCE) rather than between them (e.g., DelPhiPKa/NEQ). Strong correlations were particularly evident within the PB-based approaches when evaluating on both the FEP+ data set and the full data set (Figures 5a and S9).

Probing the full data set revealed a general decrease in the AUE and stronger correlations with experiment (Figure S9). Given that the FEP+ data set contains a higher proportion of

https://doi.org/10.1021/acs.jctc.3c00721
J. Chem. Theory Comput. 2023, 19, 7833−7845

glutamates to aspartates and no lysines, this result suggests that data set composition can impact performance and should warrant consideration in future benchmarks.

Both MD-based methods, NEQ and FEP+, showed high levels of agreement with each other and with experiment. The rather weak intermethod correlation is further emphasized by comparing consensus results from the method families over the full data set (Figure 5c).

Comparison with a null model revealed stronger correlations over the FEP+ and full data sets for all methods considered (Figures 5a and S9). However, the average unsigned errors for several approaches were not significantly different from the errors of the null model. The MD-based approach exhibited consistent performance even for residues with $|\Delta pK_a| > 1$ (Figure S3), performing significantly better than the null model, where the AUE degrades linearly with ΔpK_a.

Overall the MD-based approach was the only method to match or significantly exceed the null model with respect to the average unsigned error and Pearson correlation coefficient across all three residue classes (Figure S10). Among the predictors, both PropKa and PypKa performed well on the FEP+ and full data sets; with the exception of pK_a-ANI, these represent the two strongest performing, non-MD methods evaluated here.

Machine Learning Predictor pK_a-ANI. We also evaluated the performance of a promising, recently developed machine-llearning-based predictor, pK_a-ANI. Unfortunately, the set of pK_a values collected in this work largely overlapped with the training set of pK_a-ANI. As the evaluation of an ML approach on its training set should not be used to judge the accuracy of the method, we present this evaluation only for the sake of completeness (Figure S8). As expected, pK_a-ANI performance on the full data set was strong, exceeding the other methods with respect to AUE (0.44 ± 0.07 pK) and Pearson correlation coefficient (0.87 ± 0.05).

To gain a more realistic insight into the performance of pK_a-ANI, we considered a small subset of 14 pK_a values from the full data set that did not appear in the training set of pK_a-ANI. This set, however, contains only lysine residues from two protein systems. The observed accuracy on this subset was 0.49 ± 0.10 pK with a correlation of −0.18 ± 0.27 (compared to 0.48 ± 0.07 and 0.71 ± 0.16 with the MD-based approach) (Figure 6).

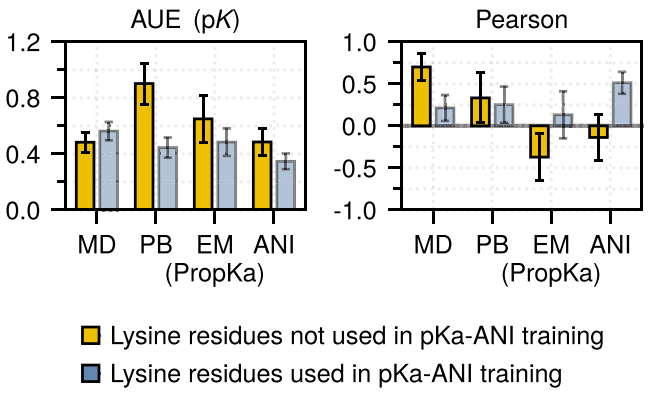


Figure 6. Performance of methods on the lysine subset (14 values), which were not in the pK_a-ANI training set. The performance on the rest of the lysine set (25 values) is shown as a reference. Bootstrapped standard errors are depicted.

We can assess the accuracy difference between the "train" and "test" sets by evaluating the performance of pK_a-ANI on a lysine pK_a subset that was used to train the predictor.

While in terms of AUE the performance of the ML-based predictor becomes only insignificantly worse, the reduction in the Pearson correlation coefficient between the "test" subset and the "training" set is significant. Given the small size of the "test" data set and bias toward only one residue type, this evaluation of pK_a-ANI accuracy should not be overinterpreted. Nevertheless, our analysis suggests a reduction in prediction accuracy when using independent test data, a result consistent with the original pK_a-ANI publication.[70]

DISCUSSION

Here, we assess the ability of NEQ-based free energy calculations to resolve the pK_a values of 144 residues across 13 proteins. Although large-scale studies on the application of NEQ alchemical calculations for predicting mutagenic folding free energy changes and relative and absolute ligand-binding affinities already exist, such an extension to protein pK_a values has been absent from the literature. A seamless free energy workflow that can probe the role of protonation on ligand binding, particularly relevant at an enzymatic active site, and resolve the underlying pK_a values of both individual residues and bound molecules is highly desirable. Here, we take a step toward that goal. Although (de)protonation is the smallest topological change that a residue can undergo, it results in a significant charge shift. We find that such perturbations and the corresponding free energies can be readily resolved using our `pmx`-based approach (i.e., AUE: 0.68 ± 0.05 pK), with accuracy comparable to FEP+,[58] and demonstrate the ability of this approach to resolve the pK_a of coupled residues. While the MD-based approach can capture protein dynamics and account for residue coupling, with both contributing to the accurate pK_a predictions, it is a computationally expensive method. Based on the timings from the current work, running simulations for 1 week on a single GPU (RTX 2080 Ti) would allow for computing 12 pK_a differences in an average-sized protein domain (≈100 residues).

Our results reveal that the Amber14SB[74]/Amber99SB*-ILDN[142] partial charges for lysine are likely erroneous, yielding pK_a and thermostability estimates that deviate significantly from experiment. Importantly, we demonstrate that this error can be resolved using charges assigned in Amber99SB*-ILDN-Q.[75] Taken alongside those by Best et al., our results do suggest that the Amber14SB backbone partial charges warrant further investigation; however, we do not advocate the use of Amber14SB-K until further validation is performed. One interesting point of investigation could be determining whether these modified charges resolve previously documented ion-overbinding problems[135] and conformational discrepancies in polyelectrolytes.[145]

While our results and the recent work of others[58,146] underscore the pK_a prediction accuracy attainable by MD-based free energy methods, the gap between prediction and experiment remains larger than the experimental error of 0.1−0.2 pK units.[36] In the current work, we have identified two main sources contributing to the pK_a prediction error: residue coupling and force-field parametrization.

With respect to the first, we have demonstrated that accounting for the coupling of nearby titratable sites plays a crucial role in accurate pK_a prediction. While this requires additional calculations within the alchemical free energy

https://doi.org/10.1021/acs.jctc.3c00721
J. Chem. Theory Comput. 2023, 19, 7833−7845

framework,[43] it brings a significant improvement to the prediction accuracy (Figure 3).

Regarding the second, we found that the deprotonated lysine backbone partial charges in Amber14SB are more favorable relative to the protonated backbone charges, which, in turn, results in a pK_a underestimation. In support of this hypothesis was the observation that the effect was largest for residues situated in regions where backbone interactions are most prominent (e.g., α-helix). Our finding underscores the importance of accurately parametrizing both the protonated and deprotonated forms of the amino acids and the sensitivity that relative free energy calculations can have to seemingly minor parametrization differences. Suggestive of this phenomenon was the recent demonstration[147] that modification of the Amber14SB cysteine thiolate parameters—to agree more closely with ab initio solvation data—could improve the pK_a prediction accuracy by 0.5 pK units when combined with an MD-based approach.[146] The use of polarizable force fields might also improve pK_a estimates;[148] however, recent work using Monte Carlo simulations with the Drude force field and a Poisson−Boltzmann continuum solvent model did not show a significantly improved prediction accuracy.[149]

We note that conformational sampling may also play a role; however, this is less significant in the systems probed here. For proteins with more pronounced pH-dependent conformation shifts, local rearrangements over tens of nanoseconds may be insufficient to capture the end-state distributions and would result in poorer estimates of the pK_a.[150,151]

In summary, we have shown that our open-source, `pmx`-based NEQ free energy method performs on par with state-of-the-art commercial software and achieves an average unsigned error that meets or exceeds alternative in silico predictors when assessed on independent test data. Furthermore, this MD-based approach yielded markedly stronger correlations with experiment, suggesting better performance for the discrimination of residues with similar pK_as. Additionally, our observation of a significant partial charge discrepancy suggests that high-quality experimental pK_a values may constitute a compelling data set to be used during force-field parametrization.

■ ASSOCIATED CONTENT

Data Availability Statement

PDB structures, simulation setup files, and calculated pK_a values are available at https://github.com/deGrootLab/pka_prediction_2023.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c00721.

Assessment of the simulation protocol (Figure S1); protein-wise AUEs and Pearson correlation coefficients (Figure S2); AUE for individual methods is shown as a function of ΔpK_a (Figure S3); AUE for the two force fields used as a function of solvation number (Figure S4); solvation number as a function of ΔpK_a interval (Figure S5); AUEs comparison between Amber14SB and Amber14SB-K for specific secondary structure elements (Figure S6); residue-wise AUEs and Pearson correlation coefficients computed on an aspartate/glutamate reduced subset for Amber99SB-*disp* (Figure S7); residue-wise AUEs and Pearson correlation coefficients for individual methods (Figure S8); Pearson correlation coefficients and deviations (AUEs) of the ΔpK_a between methods (Figure S9); residue-wise *p*-value analysis for the Pearson correlation coefficients and deviations (AUEs) of the ΔpK_a between methods (Figure S10); experimental conditions and pK_a values (Table S1); and experimental conditions and thermostability values (Table S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Vytautas Gapsys − *Computational Biomolecular Dynamics Group, Department of Theoretical and Computational Biophysics, Max Planck Institute for Multidisciplinary Sciences, 37077 Göttingen, Germany; Computational Chemistry, Janssen Research & Development, Janssen Pharmaceutica N. V., B-2340 Beerse, Belgium;* orcid.org/0000-0002-6761-7780; Email: vgapsys@gwdg.de

Authors

Carter J. Wilson − *Department of Mathematics, The University of Western Ontario, N6A 5B7 London, Canada; Centre for Advanced Materials and Biomaterials Research (CAMBR), The University of Western Ontario, N6A 5B7 London, Canada;* orcid.org/0000-0002-8992-6269

Mikko Karttunen − *Department of Physics & Astronomy, Department of Chemistry, and Centre for Advanced Materials and Biomaterials Research (CAMBR), The University of Western Ontario, N6A 5B7 London, Canada;* orcid.org/0000-0002-8626-3033

Bert L. de Groot − *Computational Biomolecular Dynamics Group, Department of Theoretical and Computational Biophysics, Max Planck Institute for Multidisciplinary Sciences, 37077 Göttingen, Germany;* orcid.org/0000-0003-3570-3534

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00721

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

C.J.W. thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Government of Ontario for funding. M.K. thanks the Discovery and Canada Research Chairs Program of the Natural Sciences and Engineering Research Council of Canada (NSERC) for financial support. The authors are grateful to Hatice Gokcan and Olexandr Isayev for their comments on the use of pKa-ANI.

■ REFERENCES

(1) The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506−D515, DOI: 10.1093/nar/gky1049.

(2) Jordan, I. K.; Kondrashov, F. A.; Adzhubei, I. A.; Wolf, Y. I.; Koonin, E. V.; Kondrashov, A. S.; Sunyaev, S. A universal trend of amino acid gain and loss in protein evolution. *Nature* **2005**, *433*, 633−638.

(3) Yang, A.-S.; Honig, B. On the pH Dependence of Protein Stability. *J. Mol. Biol.* **1993**, *231*, 459−474.

(4) Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. Protein Ionizable Groups: pK Values and Their Contribution to Protein Stability and Solubility. *J. Biol. Chem.* **2009**, *284*, 13285−13289.
(5) Schaefer, M.; Sommer, M.; Karplus, M. pH-Dependence of Protein Stability: Absolute Electrostatic Free Energy Differences between Conformations. *J. Phys. Chem. B* **1997**, *101*, 1663−1683.
(6) Tollinger, M.; Crowhurst, K. A.; Kay, L. E.; Forman-Kay, J. D. Site-specific contributions to the pH dependence of protein stability. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 4545−4550.
(7) Shaw, K. L.; Grimsley, G. R.; Yakovlev, G. I.; Makarov, A. A.; Pace, C. N. The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. *Protein Sci.* **2001**, *10*, 1206−1215.
(8) Kramer, R. M.; Shende, V. R.; Motl, N.; Pace, C. N.; Scholtz, J. M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys. J.* **2012**, *102*, 1907−1915.
(9) Watanabe, H.; Yoshida, C.; Ooishi, A.; Nakai, Y.; Ueda, M.; Isobe, Y.; Honda, S. Histidine-Mediated Intramolecular Electrostatic Repulsion for Controlling pH-Dependent Protein−Protein Interaction. *ACS Chem. Biol.* **2019**, *14*, 2729−2736.
(10) Sheinerman, F. B.; Norel, R.; Honig, B. Electrostatic aspects of protein−protein interactions. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153−159.
(11) Paulsen, C. E.; Carroll, K. S. Cysteine-Mediated Redox Signaling: Chemistry, Biology, and Tools for Discovery. *Chem. Rev.* **2013**, *113*, 4633−4679.
(12) Isom, D. G.; Dohlman, H. G. Buried ionizable networks are an ancient hallmark of G protein-coupled receptor activation. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 5702−5707.
(13) Dwyer, J. J.; Gittis, A. G.; Karp, D. A.; Lattman, E. E.; Spencer, D. S.; Stites, W. E.; García-Moreno, E. B. High Apparent Dielectric Constants in the Interior of a Protein Reflect Water Penetration. *Biophys. J.* **2000**, *79*, 1610−1620.
(14) Harms, M. J.; Castañeda, C. A.; Schlessman, J. L.; Sue, G. R.; Isom, D. G.; Cannon, B. R.; García-Moreno, E. B. The pKa Values of Acidic and Basic Residues Buried at the Same Internal Location in a Protein Are Governed by Different Factors. *J. Mol. Biol.* **2009**, *389*, 34−47.
(15) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; Velu, P. D.; García-Moreno, E. B. Charges in the hydrophobic interior of proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 16096−16100.
(16) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; García-Moreno, E. B. Large shifts in pKa values of lysine residues buried inside a protein. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 5260−5265.
(17) Stites, W. E.; Gittis, A. G.; Lattman, E. E.; Shortle, D. In a staphylococcal nuclease mutant the side-chain of a lysine replacing valine 66 is fully buried in the hydrophobic core. *J. Mol. Biol.* **1991**, *221*, 7−14, DOI: 10.1016/0022-2836(91)80195-z.
(18) Zhang, M.; Vogel, H. Determination of the side chain pKa values of the lysine residues in calmodulin. *J. Biol. Chem.* **1993**, *268*, 22420−22428.
(19) Thompson, J. E.; Raines, R. T. Value of General Acid-Base Catalysis to Ribonuclease A. *J. Am. Chem. Soc.* **1994**, *116*, 5467−5468.
(20) Walsh, K. A.; Neurath, H. Trypsinogen and Chymotrpsinogen as Homologous Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1964**, *52*, 884−889.
(21) Dodson, G. Catalytic triads and their relatives. *Trends Biochem. Sci.* **1998**, *23*, 347−352.
(22) Matthews, B. W.; Sigler, P. B.; Henderson, R.; Blow, D. M. Three-dimensional Structure of Tosyl-α-chymotrypsin. *Nature* **1967**, *214*, 652−656.
(23) Onufriev, A. V.; Alexov, E. Protonation and pK changes in protein−ligand binding. *Q. Rev. Biophys.* **2013**, *46*, 181−209.
(24) Kim, M. O.; Blachly, P. G.; McCammon, J. A. Conformational Dynamics and Binding Free Energies of Inhibitors of BACE-1: From the Perspective of Protonation Equilibria. *PLOS Comput. Biol.* **2015**, *11*, No. e1004341, DOI: 10.1371/journal.pcbi.1004341.
(25) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.
(26) Smith, R.; Brereton, I. M.; Chai, R. Y.; Kent, S. B. Ionization states of the catalytic residues in HIV-1 protease. *Nat. Struct. Mol. Biol.* **1996**, *3*, 946−950.
(27) Yamazaki, T.; Nicholson, L. K.; Wingfield, P.; Stahl, S. J.; Kaufman, J. D.; Eyermann, C. J.; Hodge, C. N.; Lam, P. Y. S.; Torchia, D. A. NMR and X-ray Evidence That the HIV Protease Catalytic Aspartyl Groups Are Protonated in the Complex Formed by the Protease and a Non-Peptide Cyclic Urea-Based Inhibitor. *J. Am. Chem. Soc.* **1994**, *116*, 10791−10792.
(28) Xie, D.; Gulnik, S.; Collins, L.; Gustchina, E.; Suvorov, L.; Erickson, J. W. Dissection of the pH Dependence of Inhibitor Binding Energetics for an Aspartic Protease: Direct Measurement of the Protonation States of the Catalytic Aspartic Acid Residues. *Biochemistry* **1997**, *36*, 16166−16172.
(29) Kim, M. O.; McCammon, J. A. Computation of pH-dependent binding free energies. *Biopolymers* **2016**, *105*, 43−49.
(30) Luo, R.; Head, M. S.; Moult, J.; Gilson, M. K. pKa Shifts in Small Molecules and HIV Protease: Electrostatics and Conformation. *J. Am. Chem. Soc.* **1998**, *120*, 6138−6146.
(31) Bastys, T.; Gapsys, V.; Doncheva, N. T.; Kaiser, R.; de Groot, B. L.; Kalinina, O. V. Consistent Prediction of Mutation Effect on Drug Binding in HIV-1 Protease Using Alchemical Calculations. *J. Chem. Theory Comput.* **2018**, *14*, 3397−3408.
(32) McGee, T. D.; Edwards, J.; Roitberg, A. E. pH-REMD Simulations Indicate That the Catalytic Aspartates of HIV-1 Protease Exist Primarily in a Monoprotonated State. *J. Phys. Chem. B* **2014**, *118*, 12577−12585.
(33) Markley, J. L. Observation of histidine residues in proteins by nuclear magnetic resonance spectroscopy. *Acc. Chem. Res.* **1975**, *8*, 70−80.
(34) Forman-Kay, J. D.; Clore, G. M.; Gronenborn, A. M. Relationship between electrostatics and redox function in human thioredoxin: characterization of pH titration shifts using two-dimensional homo- and heteronuclear NMR. *Biochemistry* **1992**, *31*, 3442−3452.
(35) Poon, D. K. Y.; Schubert, M.; Au, J.; Okon, M.; Withers, S. G.; McIntosh, L. P. Unambiguous Determination of the Ionization State of a Glycoside Hydrolase Active Site Lysine by 1H-15N Heteronuclear Correlation Spectroscopy. *J. Am. Chem. Soc.* **2006**, *128*, 15388−15389.
(36) Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O'Meara, F.; Søndergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.; Nielsen, J. E. Remeasuring HEWL pKa values by NMR spectroscopy: Methods, analysis, accuracy, and implications for theoretical pKa calculations. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 685−702.
(37) Sakurai, K.; Goto, Y. Principal component analysis of the pH-dependent conformational transitions of bovine β-lactoglobulin monitored by heteronuclear NMR. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15346−15351.
(38) Joshi, M. D.; Sidhu, G.; Pot, I.; Brayer, G. D.; Withers, S. G.; McIntosh, L. P. Hydrogen bonding and catalysis: a novel explanation for how a single amino acid substitution can change the ph optimum of a glycosidase 1 1Edited by M. F. Summers. *J. Mol. Biol.* **2000**, *299*, 255−279.
(39) Hass, M. A. S.; Jensen, M. R.; Led, J. J. Probing electric fields in proteins in solution by NMR spectroscopy. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 333−343.
(40) Reijenga, J.; van Hoof, A.; van Loon, A.; Teunissen, B. Development of Methods for the Determination of pKa Values. *Anal. Chem. Insights* **2013**, *8*, No. ACI.S12304, DOI: 10.4137/ACI.S12304.
(41) Zhang, Z. Y.; Dixon, J. E. Active site labeling of the Yersinia protein tyrosine phosphatase: The determination of the pKa of the active site cysteine and the function of the conserved histidine 402. *Biochemistry* **1993**, *32*, 9340−9345.

(42) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. Hydrogen Bonding Markedly Reduces the pK of Buried Carboxyl Groups in Proteins. *J. Mol. Biol.* **2006**, *362*, 594−604.
(43) Wilson, C. J.; de Groot, B. L.; Gapsys, V. *Resolving Coupled pH Titrations Using Non-equilibrium Free Energy Calculations*; ChemRxiv, 2023.
(44) Kirkwood, J. G. Theory of Solutions of Molecules Containing Widely Separated Charges with Special Application to Zwitterions. *J. Chem. Phys.* **1934**, *2*, 351−361.
(45) Tanford, C.; Kirkwood, J. G. Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres. *J. Am. Chem. Soc.* **1957**, *79*, 5333−5339.
(46) Bashford, D.; Karplus, M. pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry* **1990**, *29*, 10219−10225.
(47) Alexov, E.; Mehler, E. L.; Baker, N.; Baptista, A. M.; Huang, Y.; Milletti, F.; Nielsen, J. E.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the prediction of pKa values in proteins. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3260−3275.
(48) Sharp, K. A.; Honig, B. Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation. *J. Phys. Chem. A* **1990**, *94*, 7684−7692.
(49) Demchuk, E.; Wade, R. C. Improving the Continuum Dielectric Approach to Calculating pKas of Ionizable Groups in Proteins. *J. Phys. Chem. A* **1996**, *100*, 17373−17387.
(50) Alexov, E.; Gunner, M. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.* **1997**, *72*, 2075−2093.
(51) 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.
(52) Nielsen, J. E. On the evaluation and optimization of protein X-ray structures for pKa calculations. *Protein Sci.* **2003**, *12*, 313−326.
(53) Witham, S.; Talley, K.; Wang, L.; Zhang, Z.; Sarkar, S.; Gao, D.; Yang, W.; Alexov, E. Developing hybrid approaches to predict pKa values of ionizable groups. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3389−3399.
(54) Meyer, T.; Knapp, E.-W. pKa Values in Proteins Determined by Electrostatics Applied to Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **2015**, *11*, 2827−2840.
(55) Zheng, Y.; Cui, Q. Microscopic mechanisms that govern the titration response and pKa values of buried residues in staphylococcal nuclease mutants. *Proteins: Struct., Funct., Bioinf.* **2017**, *85*, 268−281.
(56) Simonson, T.; Carlsson, J.; Case, D. A. Proton Binding to Proteins: pKa Calculations with Explicit and Implicit Solvent Models. *J. Am. Chem. Soc.* **2004**, *126*, 4167−4180.
(57) Awoonor-Williams, E.; Rowley, C. N. Evaluation of Methods for the Calculation of the pKa of Cysteine Residues in Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 4662−4673.
(58) Coskun, D.; Chen, W.; Clark, A. J.; Lu, C.; Harder, E. D.; Wang, L.; Friesner, R. A.; Miller, E. B. Reliable and Accurate Prediction of Single-Residue pKa Values through Free Energy Perturbation Calculations. *J. Chem. Theory Comput.* **2022**, *18*, 7193−7204.
(59) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys.* **2002**, *117*, 4184−4200.
(60) Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 469−480.
(61) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **2004**, *25*, 2038−2048.
(62) Meng, Y.; Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Biomolecules Using a Discrete Protonation Model. *J. Chem. Theory Comput.* **2010**, *6*, 1401−1412.
(63) Kong, X.; Brooks, C. L. λ-dynamics: A new approach to free energy calculations. *J. Chem. Phys.* **1996**, *105*, 2414−2423.
(64) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738−752.
(65) Khandogin, J.; Brooks, C. L. Constant pH Molecular Dynamics with Proton Tautomerism. *Biophys. J.* **2005**, *89*, 141−157.
(66) Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmüller, H. Constant pH Molecular Dynamics in Explicit Solvent with λ-Dynamics. *J. Chem. Theory Comput.* **2011**, *7*, 1962−1978.
(67) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284−2295.
(68) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.
(69) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. Protein pKa Prediction by Tree-Based Machine Learning. *J. Chem. Theory Comput.* **2022**, *18*, 2673−2686.
(70) Gokcan, H.; Isayev, O. Prediction of protein pKa with representation learning. *Chem. Sci.* **2022**, *13*, 2462−2474.
(71) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angew. Chem., Int. Ed.* **2016**, *55*, 7364−7368.
(72) Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; van Vlijmen, H.; Tresadern, G.; de Groot, B. L. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **2020**, *11*, 1140−1152.
(73) Gapsys, V.; Yildirim, A.; Aldeghi, M.; Khalak, Y.; van der Spoel, D.; de Groot, B. L. Accurate absolute free energies for ligand−protein binding based on non-equilibrium approaches. *Commun. Chem.* **2021**, *4*, No. 61, DOI: 10.1038/s42004-021-00498-y.
(74) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(75) Best, R. B.; de Sancho, D.; Mittal, J. Residue-Specific α-Helix Propensities from Molecular Simulation. *Biophys. J.* **2012**, *102*, 1462−1467.
(76) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated protein structure and topology generation for alchemical perturbations. *J. Comput. Chem.* **2015**, *36*, 348−354.
(77) Parkin, S.; Rupp, B.; Hope, H. Structure of bovine pancreatic trypsin inhibitor at 125 K definition of carboxyl-terminal residues Gly57 and Ala58. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1996**, *52*, 18−29.
(78) Richarz, R.; Wüthrich, K. High field ^{13}C NMR studies at 90.5 MHz of the methyl groups in the basic pancreatic trypsin inhibitor. *FEBS Lett.* **1977**, *79*, 64−68.
(79) Brown, L. R.; Marco, A.; Wagner, G.; Wüthrich, K. A Study of the Lysyl Residues in the Basic Pancreatic Trypsin Inhibitor using 1H Nuclear Magnetic Resonance at 360 MHz. *Eur. J. Biochem.* **1976**, *62*, 103−107.
(80) Martin, C.; Richard, V.; Salem, M.; Hartley, R.; Mauguen, Y. Refinement and structural analysis of barnase at 1.5A resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 386−398.
(81) Oliveberg, M.; Arcus, V. L.; Fersht, A. R. pKa Values of Carboxyl Groups in the Native and Denatured States of Barnase: The pKA Values of the Denatured State Are on Average 0.4 Units Lower Than Those of Model Compounds. *Biochemistry* **1995**, *34*, 9424−9433.
(82) Boissy, G.; de La Fortelle, E.; Kahn, R.; Huet, J.-C.; Bricogne, G.; Pernollet, J.-C.; Brunie, S. Crystal structure of a fungal elicitor secreted by *Phytophthora cryptogea*, a member of a novel class of plant necrotic proteins. *Structure* **1996**, *4*, 1429−1439.

(83) Gooley, P. R.; Keniry, M. A.; Dimitrov, R. A.; Marsh, D. E.; Keizer, D. W.; Gayler, K. R.; Grant, B. R. The NMR solution structure and characterization of pH dependent chemical shifts of the β-elicitin, cryptogein. *J. Biomol. NMR* **1998**, *12*, 523−534, DOI: 10.1023/A:1008395001008.
(84) Hervø-Hansen, S.; Højgaard, C.; Johansson, K. E.; Wang, Y.; Wahni, K.; Young, D.; Messens, J.; Teilum, K.; Lindorff-Larsen, K.; Winther, J. R. Charge Interactions in a Highly Charge-Depleted Protein. *J. Am. Chem. Soc.* **2021**, *143*, 2500−2508.
(85) Castañeda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; García-Moreno, B. E. Molecular determinants of the pKa values of Asp and Glu residues in staphylococcal nuclease. *Proteins: Struct., Funct., Bioinf.* **2009**, *77*, 570−588.
(86) Skelton, N. J.; Kördel, J.; Chazin, W. J. Determination of the solution structure of apo calbindin D9k by NMR spectroscopy. *J. Mol. Biol.* **1995**, *249*, 441−462.
(87) Kesvatera, T.; Jönsson, B.; Thulin, E.; Linse, S. Measurement and Modelling of Sequence-specific pKa Values of Lysine Residues in Calbindin D9k. *J. Mol. Biol.* **1996**, *259*, 828−839.
(88) onu Kesvatera, T.; Jönsson, B.; Thulin, E.; Linse, S. Ionization Behavior of Acidic Residues in Calbindin D9k. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 106−115, DOI: 10.1002/(sici)1097-0134(19991001)37:1<106::aid-prot10>3.0.co;2-m.
(89) Sevcik, J.; Dauter, Z.; Lamzin, V. S.; Wilson, K. S. Ribonuclease from Streptomyces aureofaciens at Atomic Resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1996**, *52*, 327−344.
(90) Laurents, D. V.; Huyghues-Despointes, B. M.; Bruix, M.; Thurlkill, R. L.; Schell, D.; Newsom, S.; Grimsley, G. R.; Shaw, K. L.; Treviño, S.; Rico, M.; Briggs, J. M.; Antosiewicz, J. M.; Scholtz, J.; Pace, C. Charge−Charge Interactions are Key Determinants of the pK Values of Ionizable Groups in Ribonuclease Sa (pI = 3.5) and a Basic Variant (pI = 10.2). *J. Mol. Biol.* **2003**, *325*, 1077−1092.
(91) Ramanadham, M.; Sieker, L. C.; Jensen, L. H. Refinement of triclinic lysozyme: II. The method of stereochemically restrained least squares. *Acta Crystallogr., Sect. B: Struct. Sci.* **1990**, *46*, 63−69.
(92) Forman-Kay, J. D.; Clore, G. M.; Wingfield, P. T.; Gronenborn, A. M. High-resolution three-dimensional structure of reduced recombinant human thioredoxin in solution. *Biochemistry* **1991**, *30*, 2685−2698.
(93) Qin, J.; Clore, G. M.; Gronenborn, A. M. Ionization Equilibria for Side-Chain Carboxyl Groups in Oxidized and Reduced Human Thioredoxin and in the Complex with Its Target Peptide from the Transcription Factor NFκB. *Biochemistry* **1996**, *35*, 7−13.
(94) Katayanagi, K.; Miyagawa, M.; Matsushima, M.; Ishikawa, M.; Kanaya, S.; Nakamura, H.; Ikehara, M.; Matsuzaki, T.; Morikawa, K. Structural details of ribonuclease H from *Escherichia coli* as refined to an atomic resolution. *J. Mol. Biol.* **1992**, *223*, 1029−1052.
(95) Oda, Y.; Yamazaki, T.; Nagayama, K.; Kanaya, S.; Kuroda, Y.; Nakamura, H. Individual Ionization Constants of All the Carboxyl Groups in Ribonuclease HI from *Escherichia coli* Determined by NMR. *Biochemistry* **1994**, *33*, 5275−5284.
(96) Hoogstraten, C. G.; Choe, S.; Westler, W. M.; Markley, J. L. Comparison of the accuracy of protein solution structures derived from conventional and network-edited NOESY data. *Protein Sci.* **1995**, *4*, 2289−2299.
(97) Schaller, W.; Robertson, A. D. pH, Ionic Strength, and Temperature Dependences of Ionization Equilibria for the Carboxyl Groups in Turkey ovomucoid Third Domain. *Biochemistry* **1995**, *34*, 4714−4723.
(98) DeRose, E. F.; Kirby, T. W.; Mueller, G. A.; Bebenek, K.; Garcia-Diaz, M.; Blanco, L.; Kunkel, T. A.; London, R. E. Solution Structure of the Lyase Domain of Human DNA Polymerase λ. *Biochemistry* **2003**, *42*, 9564−9574.
(99) Gao, G.; DeRose, E. F.; Kirby, T. W.; London, R. E. NMR Determination of Lysine pKa Values in the Pol λ Lyase Domain: Mechanistic Implications. *Biochemistry* **2006**, *45*, 1785−1794.
(100) Ishida, H.; Nakashima, K.-i.; Kumaki, Y.; Nakata, M.; Hikichi, K.; Yazawa, M. The Solution Structure of Apocalmodulin from *Saccharomyces cerevisiae* Implies a Mechanism for Its Unique Ca2+ Binding Property. *Biochemistry* **2002**, *41*, 15536−15542, DOI: 10.1021/bi020330r.
(101) Chen, J.; Lu, Z.; Sakon, J.; Stites, W. E. Increasing the thermostability of staphylococcal nuclease: implications for the origin of protein thermostability. *J. Mol. Biol.* **2000**, *303*, 125−130.
(102) Eftink, M. R.; Ghiron, C. A.; Kautz, R. A.; Fox, R. O. Fluorescence and conformational stability studies of Staphylococcus nuclease and its mutants, including the less stable nuclease-concanavalin A hybrids. *Biochemistry* **1991**, *30*, 1193−1199.
(103) Weaver, L.; Matthews, B. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *J. Mol. Biol.* **1987**, *193*, 189−199.
(104) Heinz, D. W.; Baase, W. A.; Matthews, B. W. Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 3751−3755.
(105) Blaber, M.; Zhang, X.-j.; Lindstrom, J. D.; Pepiot, S. D.; Baase, W. A.; Matthews, B. W. Determination of α-Helix Propensity within the Context of a Folded Protein. *J. Mol. Biol.* **1994**, *235*, 600−624.
(106) Lipscomb, L. A.; Gassner, N. C.; Snow, S. D.; Eldridge, A. M.; Baase, W. A.; Drew, D. L.; Matthews, B. W. Context-dependent protein stabilization by methionine-to-leucine substitution shown in T4 lysozyme. *Protein Sci.* **1998**, *7*, 765−773.
(107) Dao-pin, S.; Söderlind, E.; Baase, W. A.; Wozniak, J. A.; Sauer, U.; Matthews, B. W. Cumulative site-directed charge-change replacements in bacteriophage T4 lysozyme suggest that long-range electrostatic interactions contribute little to protein stability. *J. Mol. Biol.* **1991**, *221*, 873−887.
(108) Nicholson, H.; Tronrud, D. E.; Becktel, W. J.; Matthews, B. W. Analysis of the effectiveness of proline substitutions and glycine replacements in increasing the stability of phage T4 lysozyme. *Biopolymers* **1992**, *32*, 1431−1441.
(109) Mooers, B. H. M.; Baase, W. A.; Wray, J. W.; Matthews, B. W. Contributions of all 20 amino acids at site 96 to the stability and structure of T4 lysozyme. *Protein Sci.* **2009**, *18*, 871−880.
(110) Nicholson, H.; Söderlind, E.; Tronrud, D.; Matthews, B. Contributions of left-handed helical residues to the structure and stability of bacteriophage T4 lysozyme. *J. Mol. Biol.* **1989**, *210*, 181−193.
(111) Ishikawa, K.; Kimura, S.; Kanaya, S.; Morikawa, K.; Nakamura, H. Structural study of mutants of *Escherichia coli* ribonuclease HI with enhanced thermostability. *Protein Eng., Des. Sel.* **1993**, *6*, 85−91.
(112) Gapsys, V.; Michielssens, S.; Peters, J. H.; de Groot, B. L.; Leonov, H.Calculation of Binding Free Energies. In *Molecular Modeling of Proteins*; Kukol, A., Ed.; Methods in Molecular Biology; Humana Press: New York, NY, 2015; Vol. *1215*, pp 173−209.
(113) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71−73.
(114) 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.; Wiórkiewicz-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.
(115) 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.
(116) van Gunsteren, W. F.; Berendsen, H. J. C. A Leap-frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1988**, *1*, 173−185.
(117) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(118) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(119) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 116−122.
(120) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245−268.
(121) Crooks, G. E. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E* **1999**, *60*, 2721−2726.
(122) Gapsys, V.; de Groot, B. L. On the importance of statistics in molecular simulations for thermodynamics, kinetics and simulation box size. *eLife* **2020**, *9*, No. e57589, DOI: 10.7554/eLife.57589.
(123) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK values of the ionizable groups of proteins. *Protein Sci.* **2006**, *15*, 1214−1218.
(124) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A summary of the measured pKa values of the ionizable groups in folded proteins. *Protein Sci.* **2008**, *18*, 247−251, DOI: 10.1002/pro.19.
(125) Rocchia, W.; Alexov, E.; Honig, B. Extending the Applicability of the Nonlinear Poisson-Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. *J. Phys. Chem. B* **2001**, *105*, 6507−6514.
(126) Wang, L.; Li, L.; Alexov, E. pKa predictions for proteins, RNAs, and DNAs with the Gaussian dielectric function using DelPhipKa. *Proteins: Struct., Funct., Bioinf.* **2015**, *83*, 2186−2197.
(127) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: automating pKa prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012**, *40*, W537−W541.
(128) Song, Y.; Mao, J.; Gunner, M. R. MCCE2: Improving protein pKa calculations with extensive side chain rotamer sampling. *J. Comput. Chem.* **2009**, *30*, 2231−2247.
(129) Reis, P. B. P. S.; Vila-Viçosa, D.; Rocchia, W.; Machuqueiro, M. PypKa: A Flexible Python Module for Poisson−Boltzmann-Based pKa Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 4442−4448.
(130) Bashford, D. An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules An Experience Report on the MEAD Project. In *Scientific Computing in Object-Oriented Parallel Environments. ISCOPE 1997*; Ishikawa, Y.et al., Ed.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 1997; pp 233−240.
(131) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.
(132) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192−4202.
(133) Catte, A.; Girych, M.; Javanainen, M.; Loison, C.; Melcr, J.; Miettinen, M. S.; Monticelli, L.; Määttä, J.; Oganesyan, V. S.; Ollila, O. H. S.; Tynkkynen, J.; Vilov, S. Molecular electrometer and binding of cations to phospholipid bilayers. *Phys. Chem. Chem. Phys.* **2016**, *18*, 32560−32569.
(134) Yoo, J.; Aksimentiev, A. New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Phys. Chem. Chem. Phys.* **2018**, *20*, 8432−8449.
(135) Tolmachev, D. A.; Boyko, O. S.; Lukasheva, N. V.; Martinez-Seara, H.; Karttunen, M. Overbinding and Qualitative and Quantitative Changes Caused by Simple Na^+ and K^+ Ions in Polyelectrolyte Simulations: Comparison of Force Fields with and without NBFIX and ECC Corrections. *J. Chem. Theory Comput.* **2020**, *16*, 677−687.
(136) Florová, P.; Sklenovský, P.; Banáš, P.; Otyepka, M. Explicit Water Models Affect the Specific Solvation and Dynamics of Unfolded Peptides While the Conformational Behavior and Flexibility of Folded Peptides Remain Intact. *J. Chem. Theory Comput.* **2010**, *6*, 3569−3579.
(137) Åqvist, J. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem. A* **1990**, *94*, 8021−8024.
(138) Joung, I. S.; Cheatham, 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.
(139) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **2015**, *119*, 5113−5123.
(140) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712−725.
(141) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.
(142) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950−1958.
(143) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(144) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E4758 DOI: 10.1073/pnas.1800690115.
(145) Lukasheva, N.; Tolmachev, D.; Martinez-Seara, H.; Karttunen, M. Changes in the Local Conformational States Caused by Simple Na. and K+ Ions in Polyelectrolyte Simulations: Comparison of Seven Force Fields with and without NBFIX and ECC Corrections. *Polymers* **2022**, *14*, No. 252, DOI: 10.3390/polym14020252.
(146) Awoonor-Williams, E.; Golosov, A. A.; Hornak, V. Benchmarking pKa Tools for Cysteine pKa Prediction. *J. Chem. Inf. Model.* **2023**, *63*, 2170−2180.
(147) Pedron, F. N.; Messias, A.; Zeida, A.; Roitberg, A. E.; Estrin, D. A. Novel Lennard-Jones Parameters for Cysteine and Selenocysteine in the AMBER Force Field. *J. Chem. Inf. Model.* **2023**, *63*, 595−604.
(148) Kaminski, G. A. Accurate Prediction of Absolute Acidity Constants in Water with a Polarizable Force Field: Substituted Phenols, Methanol, and Imidazole. *J. Phys. Chem. B* **2005**, *109*, 5884−5890.
(149) Aleksandrov, A.; Roux, B.; MacKerell, A. D. pKa calculations with the Polarizable Drude Force Field and Poisson−Boltzmann Solvation Model. *J. Chem. Theory Comput.* **2020**, *16*, 4655−4668.
(150) Sarkar, A.; Roitberg, A. E. pH-Dependent Conformational Changes Lead to a Highly Shifted pKa for a Buried Glutamic Acid Mutant of SNase. *J. Phys. Chem. B* **2020**, *124*, 11072−11080.
(151) Di Russo, N. V.; Estrin, D. A.; Martí, M. A.; Roitberg, A. E. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLOS Comput. Biol.* **2012**, *8*, No. e1002761, DOI: 10.1371/journal.pcbi.1002761.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on October 11, 2023, with errors in Table S1. The corrected version was reposted on October 16, 2023.

https://doi.org/10.1021/acs.jctc.3c00721
J. Chem. Theory Comput. 2023, 19, 7833−7845