Resolving coupled pH titrations using non-equilibrium free energy calculations

Carter J. Wilson,[†,‡] Bert L. de Groot,[¶] and Vytautas Gapsys[*,¶,§]

†*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*
¶*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., Turnhoutseweg 30, B-2340 Beerse, Belgium.*

E-mail: vgapsys@gwdg.de

Abstract

In a protein, nearby titratable sites can be coupled: the (de)protonation of one may affect the other. The degree of this interaction depends on several factors and can influence the measured pK_a. Here, we derive a formalism based on double free energy differences ($\Delta\Delta G$) for quantifying the individual site pK_a values of coupled residues. As $\Delta\Delta G$ values can be obtained by means of alchemical free energy calculations, the presented approach allows for a convenient estimation of coupled residue pK_as in practice. We demonstrate that our approach and a previously proposed microscopic pK_a formalism, can be combined with non-equilibrium (NEQ) alchemical free energy calculations to resolve pH-dependent protein pK_a values. Toy models and both, regular and constant-pH molecular dynamics simulations, alongside experimental data, are used to validate this approach. Our results highlight the insights gleaned when coupling and microstate probabilities are analyzed and suggest extensions to more complex enzymatic contexts. Furthermore, we find that näively computed pK_a values that ignore coupling, can be significantly improved when coupling is accounted for, in some cases reducing the error by half. In short, our results suggest that free energy methods can resolve the pK_a values of both uncoupled and coupled residues.

Introduction

Protein function is known to depend on the acidity of the medium.[1–5] Such a pH dependence is caused by the (de)protonation of amino acid residues, whereby a proton is added or removed from an amino acid side chain. As this process is pH-dependent, at certain pH levels, the event will be more or less favorable and, by definition, at the pK_a, it will be equally probable (i.e., $\Delta G_{prot} = 0$). Knowledge of the residue pK_a values in a protein is essential for understanding function. It not only allows for a rationalization of protein properties (e.g., stability,[6] solubility,[7] etc.) and interactions at a specific pH,[8,9] but in the context of enzymatic and redox reactions, pK_a values can provide insight into how favorable a proton transfer will be under certain conditions.[10]

As alluded to, the pK_a is fundamentally a free energy relationship; for an isolated, protonatable group, the value is proportional to the free energy of protonation:

$$\Delta G_{prot} = RT \log(10) (pH - pK_a) . \qquad (1)$$

This relationship suggests that the free energy is linearly dependent on the solution pH; as the pH moves farther away from pK_a, the free energy required to (de)protonate also shifts. A purely linear relationship between pH and ΔG implies a

joint relationship with the probability of finding a protonatable group i in a given state; this follows from the rearranged Henderson–Hasselbalch (HH) equation:

$$pK_a^i = pH + \log_{10}\left(\frac{\langle x_i \rangle}{1 - \langle x_i \rangle}\right), \qquad (2)$$

where $\langle x_i \rangle$ is the probability that residue i is protonated. However, such a curve, when computed from experiment, may be flatter or irregularly shaped, often necessitating the application of specialized fitting procedures.[11] In such cases, not only does the curve suggest a non-linear dependence, an analysis of the complete pH-dependent behaviour is often more insightful than defining the residue by a single pK_a value.

In proteins and, in particular, enzymes, protonatable residues can, in only a few cases, be separated from their interactions with one another.[12–14] Although these associations will be more pronounced at an active site, even more distant residues can experience some degree of coupling,[15] interacting more or less strongly depending on their microenvironment, the pH of the solution, and their own protonation state. Indeed, this could result in a more challenging resolution of "the pK_a";[11] however, such interactions may provide insight into a reaction mechanism or suggest the functional importance of a residue pair. In these scenarios, a modified HH-curve may still yield two clear inflection points; however, the assignment of pK_a values to specific residues could remain a challenge. Moreover, the protonation probability of a coupled residue, although potentially described by an HH-curve, is nonetheless a composite probability of microstates. As Edsall and Wyman[16] and later Alexey Onufriev[17] and G. Matthias Ullmann[18] helped formalize, these states are in a pH-dependent equilibrium with each other and collectively comprise the macroscopic probability observed experimentally. This knowledge gap between the measurable macrostates and the cryptic microstates suggests a potential role for theoretical and computational methods, which may help to resolve both the macroscopic pK_a and the microscopic pK_a values; we consider one of these methods here.

In summary, titration curves and pK_a values may exhibit diverse pH-dependent behaviors due to the coupling of titratable sites. The pK_as of such coupled site residues may be difficult to resolve, and even if a curve is resolved, a singular pK_a may overlook unique functionally relevant microstates. Probing these states and the microscopic pK_a values between them using free energy calculations based on a rigorous formalism may be a worthwhile approach to provide additional insight into this key biophysical phenomenon. In this work, we derive a formalism to quantify individual site pK_as in coupled residues starting from double free energy differences. This is particularly convenient, as such ΔΔG values can be efficiently computed by means of alchemical free energy calculations.

We begin by considering the relationship between free energies and pK_a values, introduce our thermodynamic cycle-based formalism, and outline the concept of microscopic pK_a values. We then demonstrate the applicability of these concepts to non-equilibrium, alchemical pK_a calculations, and show how these can provide insight beyond what could be obtained from a näive approach without taking into account residue couplings.

Theory

pK_a values and free energies

Consider the thermodynamic cycle given in Figure 1. To calculate the absolute protein pK_a value of a single residue (A), we must consider the free energy of proton transfer from the gas (g) phase into the solution (s) phase and then from the solution into the protein (p) phase. However, for many model compounds, the free energy associated with the proton transfer in solution is known. Using this reference pK_a value (pK_a°) allows us to only consider the free energies associated with the rightmost cycle, thus reducing our problem to solving $\Delta\Delta G_{s,p}(A^H, A^-)$ — the free energy associated with moving the charge from the solution site to the protein site — which is related to the protein pK_a by

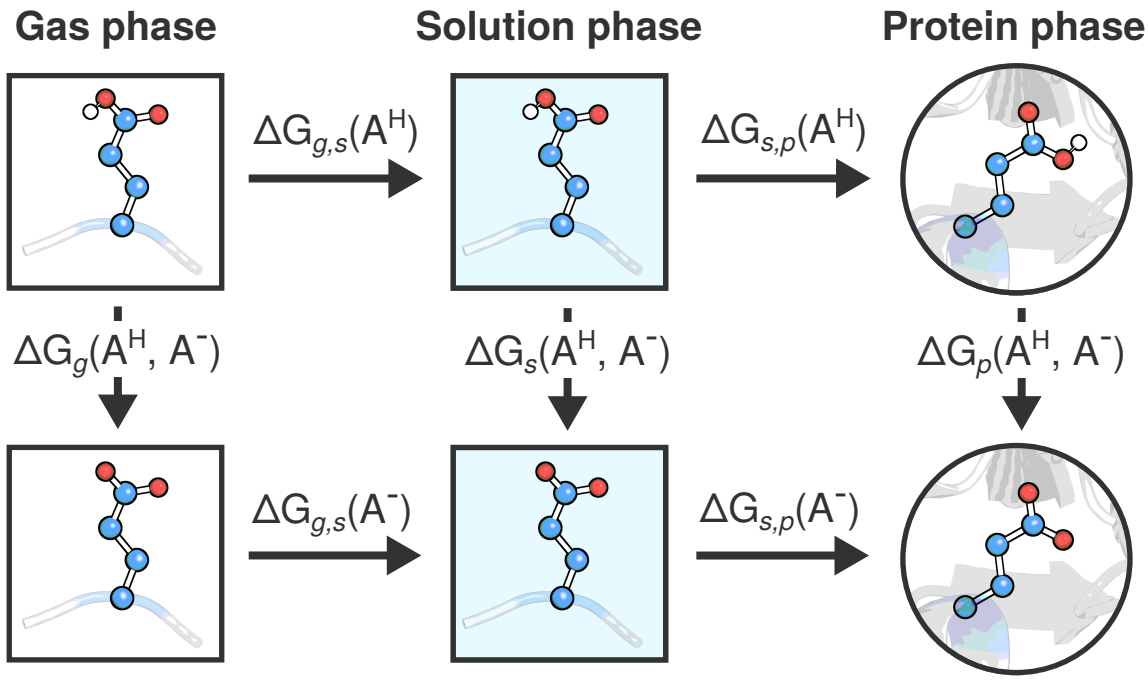

Figure 1: Complete pK_a thermodynamic cycle. The horizontal arrows mark the transfer of a titratable residue (A) between different environments: gas (g), solution (s), protein (p). The vertical arrows denote the free energy difference between the deprotonated and protonated form in a corresponding environment.

$$\begin{aligned} \mathrm{p}K_a(\text{protein}) &= \mathrm{p}K_a^\circ + \frac{\Delta G_{s,p}(A^-) - \Delta G_{s,p}(A^H)}{RT \log(10)} \\ &= \mathrm{p}K_a^\circ + \frac{\Delta\Delta G_{s,p}(A^H, A^-)}{RT \log(10)}. \end{aligned} \tag{3}$$

Equation 3 implicitly contains two terms, which we here call $\Delta\Delta G^{\text{env}}_{s,p}$ and $\Delta G^{\text{titr}}_{s,p}$. The first ($\Delta\Delta G^{\text{env}}_{s,p}$) represents the free energy of dissociating a proton within a protein relative to the solvent environment, which we represent by a capped peptide. It is assumed that the protein is fixed in some protonation state and, based on Tanford and Kirkwood,[19] has been taken to be the state in which all titratable sites in the protein are neutralized. Because these sites are fixed to their neutral state, this free energy is pH-independent. The second free energy component ($\Delta G^{\text{titr}}_{s,p}$) reintroduces pH dependence by capturing how the free energy of dissociating the proton within the protein will be more or less favorable, depending on the states of the other protonatable sites.

The contribution of the free energy component $\Delta\Delta G^{\text{env}}_{s,p}$ is determined by the difference in the solvation free energies of the species (i.e., $\Delta\Delta G^{\text{env}}_{s,p} = \Delta G^{\text{env}}_p - \Delta G^{\text{env}}_s$). In the case of the peptide (ΔG^{env}_s), the contribution is governed almost entirely by the solvent; however, in the protein (ΔG^{env}_p), van der Waals and electrostatic interactions with permanent dipoles dominate.[20] Note that while the desolvation penalty of moving the proton out of the solution and into the protein may be large — depending on the solvent exposure of the residue in the protein — favorable interactions between neighboring residues can compensate for this, stabilizing the buried residue in its charged or neutral state.

Once a reference protonation state is set, the $\Delta\Delta G^{\text{env}}_{s,p}$ can be resolved and an intrinsic pK_a (pK_{int}) can be defined:

$$\mathrm{p}K_{\text{int}} = \mathrm{p}K_a^\circ + \frac{\Delta\Delta G^{\text{env}}_{s,p}}{RT \log(10)}. \tag{4}$$

Again, note that this is pH-independent as all other residues are in a fixed protonation state without the ability to titrate. The true pK_a for a residue in a protein will depend on the dynamic protonation state of these other residues and will have a pH

dependence:

$$\mathrm{p}K_\mathrm{a}(\mathrm{protein}) = \mathrm{p}K_\mathrm{int} + \frac{\Delta \mathrm{G}^\mathrm{titr}_{s,p}(\mathrm{pH})}{\mathrm{RT}\log(10)}. \quad (5)$$

Although the $\mathrm{p}K_\mathrm{int}$ is pH-independent, it may still provide a strong estimate of the true $\mathrm{p}K_\mathrm{a}$ depending on the reference protonation state assigned to the protein. Nevertheless, the assumption that $\mathrm{p}K_\mathrm{a} \approx \mathrm{p}K_\mathrm{int}$ will fail in some instances and only by considering $\Delta \mathrm{G}^\mathrm{titr}_{s,p}(\mathrm{pH})$ can an accurate $\mathrm{p}K_\mathrm{a}$(protein) be resolved.

If all protonatable residues were allowed to titrate, then computing $\Delta \mathrm{G}^\mathrm{titr}_{s,p}(\mathrm{pH})$ would require a consideration of all pairs of relevant interactions. Inevitably, most of these will contribute very little to this energy, potentially resulting in frivolous calculations. Instead, one might assume that all distant protonatable residues are fixed to their model states at pH 7.4 (i.e., Asp/Glu: deprotonated, Lys/Cys: protonated), and that the only relevant contribution of $\Delta \mathrm{G}^\mathrm{titr}(\mathrm{pH})$ with respect to some protonatable group A, comes from the nearest protonatable group B, given that A and B are close to each other (e.g., $\mathrm{r}_\mathrm{AB} < 0.5$ nm). Under these assumptions, one improves computational efficiency and may not sacrifice an accurate solution. Here, we will consider these two approaches for resolving the $\mathrm{p}K_\mathrm{a}$ of residue A: 1) we assume $\mathrm{p}K_\mathrm{a} = \mathrm{p}K_\mathrm{int}$ and 2) we assume $\mathrm{p}K_\mathrm{a} = \mathrm{p}K_\mathrm{int} + \frac{\Delta \mathrm{G}^\mathrm{titr}(\mathrm{pH})}{\mathrm{RT}\log(10)}$ and that only the protonatable residue closest to A is titratable. In both cases, we assume that all residues in the protein are assigned to their corresponding model states at pH 7.4.

We begin with a discussion of the application of non-equilibrium (NEQ) free energy calculations to the problem of computing the two aforementioned values, namely: $\Delta\Delta \mathrm{G}^\mathrm{env}$ and $\Delta \mathrm{G}^\mathrm{titr}(\mathrm{pH})$.

NEQ free energies and a thermodynamic cycle-based formalism

NEQ alchemical free energy calculations are particularly well suited for computing $\Delta\Delta \mathrm{G}^\mathrm{env}_{s,p}$ in Equation 3. Within the NEQ framework, fully atomistic molecular dynamics simulations are used in conjunction with thermodynamic integration to compute the non-equilibrium, alchemical work distributions associated with transforming a structure in one state into a structure in another state. The simulations are set up in such a way that within a single system, both the residue of interest (A_p), situated in a protein, and the same residue in a blocked peptide (A_s) are present. These are restrained to prevent consequential interactions, and then the work required to alchemically transform (i.e., $\mathrm{A}^\mathrm{H}_p \rightarrow \mathrm{A}^-_p$ and $\mathrm{A}^-_s \rightarrow \mathrm{A}^\mathrm{H}_s$) these residues into their complement (i.e., (de)protonated form) is computed. This construct ensures a neutral simulation box at all times during an alchemical transition.[21]

Given two equilibrium ensembles (e.g., A_p and A_s), the distributions of work values generated by rapidly transforming residues from the first ensemble into residues from the second (and vice versa) allow one to estimate the free energy difference. Here, the transformation is alchemical, the transitions are on the order of 100 ps, and the free energy difference is estimated using Bennett's acceptance ratio[22,23] (BAR) relying on the Crooks fluctuation theorem[24] (CFT). Previous work has demonstrated the ability of this NEQ approach to resolve folding free energies[25] and binding affinities[26,27] within experimental uncertainty.

$$\begin{array}{ccc} \mathrm{A}^\mathrm{H}_p + \mathrm{B}^\mathrm{H}_p & \xrightarrow{\Delta\Delta \mathrm{G}_0} & \mathrm{A}^-_p + \mathrm{B}^\mathrm{H}_p \\ \Big\downarrow \Delta\Delta \mathrm{G}_1 & & \Big\downarrow \Delta\Delta \mathrm{G}_2 \\ \mathrm{A}^\mathrm{H}_p + \mathrm{B}^-_p & \xrightarrow{\Delta\Delta \mathrm{G}_3} & \mathrm{A}^-_p + \mathrm{B}^-_p \end{array}$$

Figure 2: Scheme I: free energy cycle for a coupled residue scenario. The branches report the free energy change associated with the deprotonation of one residue, while the other is kept fixed with respect to the free energy change associated with the corresponding deprotonation in a capped peptide.

As we have mentioned in the preceding section, given the coupling between protonatable residues, $\Delta\Delta \mathrm{G}^\mathrm{env}_{s,p}$ alone may be insufficient for an accurate calculation of $\mathrm{p}K_\mathrm{a}$(protein), and the pH-dependent $\Delta \mathrm{G}^\mathrm{titr}_{s,p}$ should also be considered. To illustrate this, we consider the coupling scenario in Figure 2:

residues A and B are close together, and their protonation free energies depend on the state of the other residue.

Here, $\Delta\Delta G_0$ corresponds to $\Delta\Delta G^{env}$: the free energy of deprotonating residue A while in the presence of protonated B. Similarly, $\Delta\Delta G_3$ corresponds to the deprotonation of A in the presence of deprotonated B. In both cases, the remaining protonatable sites in the protein are fixed to their model states at pH 7.4 (i.e., Asp/Glu: deprotonated, Lys/Cys: protonated). We also have $\Delta\Delta G_1$ and $\Delta\Delta G_2$, which will shift the populations of "reactants" and "products" with respect to $\Delta\Delta G_0$ and $\Delta\Delta G_3$. Note that because of the presence of B_p, which can (de)protonate as a function of pH, this equilibrium shift depends on the pH non-linearly.

To further formalize the pK_a calculations for coupled residues, we begin by considering the thermodynamic cycle in Figure 3, where the protonation/deprotonation events are separated for the protein (p) and peptide (s) environments. In this case, we focus on the overall ΔpK_a for the deprotonation of residue A by explicitly considering all possible protonation states of residue B.

We define ΔG of the upper branch of the cycle as

$$\Delta G_{protein} = \Delta G_0 + \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_1}\right) - \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_2}\right) \quad (6)$$

and the lower branch as

$$\Delta G_{solution} = \Delta G_3 + \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_4}\right) - \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_5}\right), \quad (7)$$

with $\beta = \frac{1}{RT}$. Observing that $\Delta G_4 = \Delta G_5$ allows us to write Equation 7 as

$$\Delta G_{solution} = \Delta G_3. \quad (8)$$

Considering the definition $\Delta\Delta G = \Delta G_{protein} - \Delta G_{solution}$ we can combine Equations 6 and 8:

$$\Delta\Delta G = \Delta G_0 - \Delta G_3 + \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_1}\right) - \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_2}\right). \quad (9)$$

When considering the free energy difference between protonation in a folded protein and protonation in a capped peptide, one can resolve $\Delta\Delta G_{30} = \Delta G_0 - \Delta G_3$; however, we can also resolve ΔG_1 and ΔG_2 from $\Delta\Delta G_{41}$ and $\Delta\Delta G_{52}$, respectively. Recall that for an isolated protonatable group (e.g., capped peptide) with a known reference pK_a°, the free energy of deprotonation is linearly related to the pH via

$$\Delta G(pH) = RT\log(10)\left(pK_a^\circ - pH\right). \quad (10)$$

It follows that

$$\Delta G_1(pH) = \Delta\Delta G_{41} + \Delta G_4 = \Delta\Delta G_{41} + RT\log(10)\left(pK_a^\circ - pH\right) \quad (11)$$

and

$$\Delta G_2(pH) = \Delta\Delta G_{52} + \Delta G_5 = \Delta\Delta G_{52} + RT\log(10)\left(pK_a^\circ - pH\right). \quad (12)$$

We can substitute Equations 11 and 12 into Equation 9 and obtain $\Delta\Delta G$ as a function of pH. Note that while maintaining the correspondence, we can directly relate this back to Equations 4 and 5 by simply relabelling the components:

$$\Delta\Delta G(pH) = \Delta G_0 - \Delta G_3 + \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_1(pH)}\right) - \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_2(pH)}\right) \quad (13)$$

$$\Delta\Delta G_{s,p}(pH) = \Delta\Delta G^{env} + \Delta G^{titr}(pH) \quad (14)$$

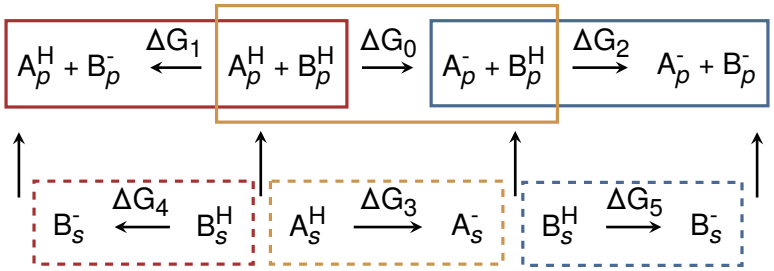

Figure 3: Scheme II: detailed free energy cycle for a coupled residue scenario. The upper branches report the free energy change associated with the deprotonation of one residue while the other is kept fixed in the protein (p), while the lower branches report the corresponding deprotonation in a capped peptide in solution (s). Note that the differences between paired upper and lower free energies are the ΔΔGs of deprotonation indicated in Figure 2.

where we see the equivalence between

$$\Delta\Delta G^{\text{env}} = \Delta G_0 - \Delta G_3, \text{ and}$$
$$\Delta G^{\text{titr}}(\text{pH}) = \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_1(\text{pH})}\right) - \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_2(\text{pH})}\right).$$

It follows that $\Delta G_{\text{protein}}$ can be expressed as a function of pH:

$$\Delta G_{\text{protein}}(\text{pH}) = \Delta\Delta G(\text{pH}) + \Delta G_3(\text{pH}). \quad (15)$$

Equation 15 provides a family of solutions that depend on the pH value. In order to determine the pK_a, we find the point where $\Delta G_{\text{protein}}(\text{pH}) = 0$; this pH corresponds to the pK_a that would be observed in a titration experiment. The result follows from the Henderson-Hasselbalch equation (Equation 2), which states that the pK_a of a residue is the pH value at which the populations of the protonated and deprotonated forms of that residue are equal (i.e., $\Delta G = 0$; the inflection point of the HH-curve). Computationally, we also have access to the whole set of pK_a solutions which are not necessarily limited by this Hesenderson-Haselbalch relation. We can combine Equations 3 and 15 and compute these pK_a values at various pH:

$$\text{p}K_a(\text{protein}) = \text{p}K_a^\circ + \frac{\Delta G_{\text{protein}}(\text{pH}) - \Delta G_3(\text{pH})}{\text{RT}\log(10)} \quad (16)$$

Moreover, instead of solving for a single pK_a value, we can also consider how the pK_a(protein) changes as a function of pH. By defining a reference state, we compute the pK_a(protein) between states and observe how the probability of microstates changes with the pH; this is the subject of consideration in the following section.

Microscopic definitions of pK_a

Complementary to the thermodynamic cycle and free energy formalisms of the preceding section is a framework based on partition functions.[18] It should be observed that the (de)protonation of the sidechain of an amino acid can be described using a standard equilibrium binding formalism. Specifically, the "binding" of protons can be fully described by the proton concentration (c) and the binding constant (K), from which it follows that the grand partition function is $\xi = 1 + K\text{c}$ and the fractional occupation of the side chain by a proton is

$$\Theta = \frac{K\text{c}}{1 + K\text{c}}. \quad (17)$$

Here, the concentration of protons in solution is related to the pH by $\text{c} = 10^{-\text{pH}}$ and the binding constant is related to the pK_a via $\text{p}K_a = -\log_{10}(K)$.

Consider a second example, involving a coupled residue system in which the protonation of one site can influence the other; specifically consider the cycle in Figure 4. Here, we have four microstates and four corresponding dissociation constants. In the fully uncoupled case, $K_1 = K_4$ and $K_2 = K_3$, and each (de)protonation event can be considered separately; this is not true when $K_1 \neq K_4$ and $K_2 \neq K_3$. In this case, we have a more complex partition

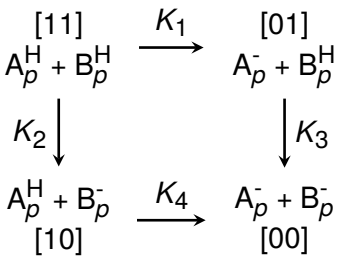

Figure 4: Two-site protonation dyad. Equilibirum constants K describe the unbinding of a proton. Values in brackets indicate the microstate (e.g., [11] : doubly protonated, [10] : first residue protonated, etc.)

function given by

$$\xi = 1 + K_1 c + K_2 c + e^{-\beta w} K_1 c K_2 c. \tag{18}$$

The form is similar to the partition function of the single-site case; however, here we include a new (un)cooperativity term which follows from the fact that: 1) the cycle is closed (i.e., $K_1 + K_3 = K_2 + K_4$), and 2) there is an "interaction free energy", w, associated with the second (de)protonation event given the first. When this interaction is zero, we have a standard two-site binding equilibrium: the proton can bind to either site, and this is governed only by the proton concentration and binding constants; however, when this interaction is positive or negative, the initial binding to one site will disfavor or favor the binding of a second proton to the other site.

This interaction notation, as defined by T. L. Hill,[28] can be related to the microstate free energy via

$$e^{-\beta w} = \frac{e^{-\beta \Delta G^\circ_{00}}}{e^{-\beta \Delta G^\circ_{01}} e^{-\beta \Delta G^\circ_{10}}}$$
$$w = \Delta G^\circ_{00} - \left(\Delta G^\circ_{01} + \Delta G^\circ_{10} \right), \tag{19}$$

where we take $\Delta G^\circ_{11} = 0$ and rely on the fact that the standard free energy of deprotonation (i.e., 1 M H^+; pH = 0) can be related to to K via

$$K = e^{-\beta \Delta G^\circ}. \tag{20}$$

When w (Equation 19) is negative, $e^{-\beta w}$ is greater than one and the unbinding of the second proton is enhanced by the first. On the contrary, when w is positive, $e^{-\beta w}$ is less than one and the second unbinding is impaired given the first. The energy of interaction, w, depends on structural changes in the protein and through-space interactions between sites. In the cases at hand, (de)protonation has only limited structural consequences and electrostatic and van der Waals were found to be dominant. Furthermore, given the repulsion of like charges, here, the unbinding of the first proton always disfavours the unbinding of the second.

Consider that we can define $\Delta G^\circ(\text{pH}) = \Delta G(0) - \mu_{H^+}$, where μ_{H^+} is the chemical potential of the protons in solution: $\mu_{H^+} = \text{RT} \log(10)\text{pH}$. We then have

$$\xi = 1 + e^{-\beta \left(\Delta G^\circ_{01}(0) - \mu_{H^+} \right)} + e^{-\beta \left(\Delta G^\circ_{10}(0) - \mu_{H^+} \right)} + e^{-\beta \left(\Delta G^\circ_{00}(0) - 2\mu_{H^+} \right)},$$

with corresponding probabilities

$$\langle A^H B^H \rangle = \frac{1}{\xi},$$
$$\langle A^- B^H \rangle = \frac{e^{-\beta \left(\Delta G^\circ_{01}(0) - \mu_{H^+} \right)}}{\xi},$$
$$\langle A^H B^- \rangle = \frac{e^{-\beta \left(\Delta G^\circ_{10}(0) - \mu_{H^+} \right)}}{\xi}, \text{ and}$$
$$\langle A^- B^- \rangle = \frac{e^{-\beta \left(\Delta G^\circ_{00}(0) - 2\mu_{H^+} \right)}}{\xi}.$$

As Ullmann notes,[18] from these we can resolve both the protonation probability of an individual site as a function of pH (e.g., probability of $A^H = \langle A^H B^H \rangle + \langle A^H B^- \rangle$), but also the probability of microstates. Observe that while in the preceding section we resolved the pK_a(protein) at a single pH for comparison with experiment, we can instead consider the family of pH-dependent solutions, substitute these into the partition function above, and thus resolve the probabilities of the individual species as a function of pH.

Methodology

Non-equilibrium alchemy

`pmx`[29] was used for the system setup, hybrid structure and topology generation, and analysis. Initial structures: Δ+PHS Staphylococcal nuclease (SNase) variant[30] (PDB: 3BDC[30]) and protein deglycase DJ-1[31] (PDB: 1P5F[32]), were taken from the PDB database. A double system in a single box setup was used, with a 3 nm distance between the protein and peptide (ACE-AXA-NH_2); this ensured charge neutrality during the alchemical transition.[21] To prevent consequential protein–peptide interactions, a single Cα in each molecule was positionally restrained. We used the CHARMM36m[33] (with CHARMM-modified TIP3P[34]) force field. A salt concentration consistent with the experimental setup was used. If no salt concentration was reported only K^+ or Cl^- counterions were added.

For all systems, an initial minimization using the steepest descent algorithm was performed. A constant temperature corresponding to the reference experimental setup was maintained implicitly using the leap-frog stochastic dynamics integrator[35] with an inverse friction constant of $\gamma = 0.5\,\mathrm{ps}^{-1}$. The pressure was maintained at 1 bar using the Parrinello-Rahman barostat[36] with a coupling time constant of 5 ps. The integration time step was set to 2 fs. Long-range electrostatic interactions were calculated using the Particle-mesh Ewald method[37] with a real-space cut-off of 1.2 nm and grid 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.[38]

To improve sampling, systems were run for 50 ns in 4 independent replicas, and the first 10 ns of each simulation was discarded as equilibration. From the remaining 40 ns, 400 non-equilibrium transitions of 500 ps each 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 difference with Bennett's acceptance ratio[22] as a maximum likelihood estimator relying on the Crooks Fluctuation Theorem.[24] Bootstrapping was used to estimate the uncertainties of the free energy estimates,[21,39] and these were propagated when calculating ΔΔG values.

Application in practice

We describe how to compute the pK_a corresponding to the upper branch in Figure 2.

1. We run three simulations using the double-system single box setup. This yields three explicit ΔΔG values (e.g., $\Delta\Delta G_0$, $\Delta\Delta G_1$, and $\Delta\Delta G_2$) and a fourth by necessity of cycle closure. These correspond to the free energy differences between deprotonation in the protein and in the capped peptide.

2. Observe that in the absence of any coupling, as the pH changes the free energy of deprotonation in the protein shifts according to

$$\Delta G(\mathrm{pH}) = \Delta\Delta G + RT \log(10)(\mathrm{p}K_a^\circ - \mathrm{pH})$$

where $\mathrm{p}K_a^\circ$ is the reference pK_a of the residue under consideration. We use this relationship to calculate

$$\Delta G_1(\mathrm{pH}) = \Delta\Delta G_1 + RT \log(10)\left(\mathrm{p}K_a^\circ - \mathrm{pH}\right)$$

and

$$\Delta G_2(\mathrm{pH}) = \Delta\Delta G_2 + RT \log(10)\left(\mathrm{p}K_a^\circ - \mathrm{pH}\right).$$

3. We can then resolve $\Delta G_{\mathrm{protein}}$ according to

$$\Delta G_{\mathrm{protein}}(\mathrm{pH}) = \Delta\Delta G(\mathrm{pH}) + \Delta G_3(\mathrm{pH})$$

where

$$\Delta\Delta G(\mathrm{pH}) = \Delta\Delta G_0 + \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_1(\mathrm{pH})}\right) - \frac{1}{\beta}\log\left(1 + e^{-\beta\Delta G_2(\mathrm{pH})}\right).$$

and

$$\Delta G_3(\mathrm{pH}) = RT \log(10)\left(\mathrm{p}K_a^\circ - \mathrm{pH}\right).$$

4. We can find the pH at which $\Delta G_{\mathrm{protein}}(\mathrm{pH}) =$

0; this pH will correspond to the apparent pK_a.

We can also compute the pK_a(protein) at arbitrary pH and construct a pH-dependent curve via

$$\mathrm{p}K_a(\mathrm{protein}) = \mathrm{p}K_a^{\circ} + \frac{\Delta G_{\mathrm{protein}}(\mathrm{pH}) - \Delta G_3(\mathrm{pH})}{RT \log(10)}.$$

(Note that if we are only interested in the pK_a, we can stop here. In order to resolve the individual site and microstate probabilities we need to follow the next three steps.)

5. Consider that we can compute standard protonation free energies via

$$\Delta G^{\circ} = \Delta G(0) - \mu_{H^+},$$

where ΔG values are from Step 2 of this protocol and μ_{H^+} is the chemical potential of the protons in the solution: $\mu_{H^+} = RT \log(10)\mathrm{pH}$.

6. We then have direct access to the partition function

$$\xi = 1 + e^{-\beta(\Delta G_0(0) - \mu_{H^+})} + e^{-\beta(\Delta G_1(0) - \mu_{H^+})} + e^{-\beta(\Delta G_0(0) + \Delta G_2(0) - 2\mu_{H^+})},$$

and corresponding microstate probabilities

$$\langle A^H B^H \rangle = \frac{1}{\xi},$$
$$\langle A^- B^H \rangle = \frac{e^{-\beta(\Delta G_0(0) - \mu_{H^+})}}{\xi},$$
$$\langle A^H B^- \rangle = \frac{e^{-\beta(\Delta G_1(0) - \mu_{H^+})}}{\xi}, \text{ and}$$
$$\langle A^- B^- \rangle = \frac{e^{-\beta(\Delta G_0(0) + \Delta G_2(0) - 2\mu_{H^+})}}{\xi}.$$

7. The overall protonation probability for an individual site can be computed from:

$$\langle A^H \rangle = \langle A^H B^H \rangle + \langle A^H B^- \rangle$$
$$\langle B^H \rangle = \langle A^H B^H \rangle + \langle A^- B^H \rangle$$

Constant-pH simulations

Constant-pH (CpHMD) simulations were performed using a recent GROMACS 2021 implementation with a modified CHARMM36m force field.[40] The system setup was similar to that discussed in the above section; however, to agree with the recommended CpHMD setup, both the leapfrog integrator and velocity rescaling with a 0.5 ps coupling time, as well as a PME Fourier spacing of 0.14 nm were used. Because only aspartate and glutamate were considered, a single-site representation (i.e., the proton can be bound to only one heavy atom in the residue) was employed; here, the A and B states represent the protonated and deprotonated forms of the titratable residue. The mass of the λ particle was set at 5 AU, and its temperature was maintained at 300 K using velocity rescaling with a 2 ps coupling time. The barrier height of the double-well potential was set at 5.0 kJ/mol. For all systems, a minimization was performed using the steepest descent algorithm followed by a 100 ns simulation in the NpT ensemble.

We considered three SNase dyads and one DJ-1 dyad, resulting in four sets of simulations. SNase simulations were performed at pH values from -1 to 7 with 0.25 pH increments and from 7 to 14 with 0.5 pH increments, while DJ-1 simulations were performed from 0 to 14 with 0.5 pH increments. Regarding SNase, both residues in the dyad (i.e., D21–D19, D21–D40, and D21–D83) were allowed to (de)protonate as a function of pH, which in two cases resulted in non-sigmoidal curves. In the case of DJ-1, we allowed only E18 to titrate, while holding C106 fixed to a protonated or deprotonated state; this resulted in sigmoidal curves.

Block averaging was used to determine the protonation probabilities and standard deviations at each pH value. Fittings to both a single and double Henderson-Hasselbalch (HH) curve were performed via bootstrap; specifically, each protonation probability was normally expanded

about its mean and a point was randomly selected from the distribution. The resultant set of points was fit to a single:

$$\langle x \rangle = \frac{10^{n(pH-pK_a)}}{1 + 10^{n(pH-pK_a)}} \quad (21)$$

and a double HH function:

$$\langle x \rangle = 0.5\frac{10^{n(pH-pK_{a1})}}{1 + 10^{n(pH-pK_{a1})}} + 0.5\frac{10^{n(pH-pK_{a2})}}{1 + 10^{n(pH-pK_{a2})}}, \quad (22)$$

and the Bayesian information criteria of each fit were used for model selection. Here, n is the Hill coefficient and controls the steepness of the titration inflection; unless otherwise stated, this is taken to be n = 1.

Results

Toy models

We first consider several systems consisting of two coupled residues (R_1, R_2), of type A, B, or C, with corresponding reference pK_a° values of 3, 4, and 8, respectively. $\Delta\Delta G$ values can be related to the dissociation constants that link microstates via Equation 5.

Model 1: Similar reference pK_a° values: no coupling

Consider the system given in Figure 5. Note

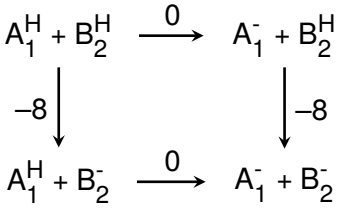

Figure 5: Thermodynamic cycle for Toy Model 1. $\Delta\Delta G$ values (kJ/mol) are indicated along the branches.

that the interaction energy between the states is zero, as evidenced by the equality of opposite paths; the protonation of A_1 has no effect on the free energy required to protonate B_2 and vice-versa. It follows that $\Delta G^{titr}(pH)$ is zero for all pH and $\Delta G_{protein}(pH)$ is constant, implying $pK_a = pK_{int}$ (Figure 7a). Computing the pK_a, we obtain values of approximately 3 and 2.6 for A_1 and B_2, respectively, unchanged and decreased from their reference values according to Equation 16.

Model 2: Similar reference pK_a° values: weak coupling

Consider the system given in Figure 6. In this

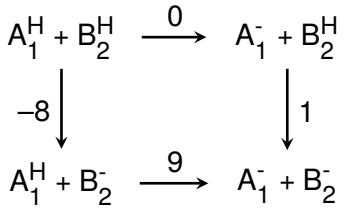

Figure 6: Thermodynamic cycle for Toy Model 2. $\Delta\Delta G$ values (kJ/mol) are indicated along the branches.

example, unlike in the first, the interaction energy is non-zero. The free energies suggest that the first deprotonation event results in a less favorable second deprotonation. This is to be expected for nearby coupled residues, where the electrostatic repulsion associated with the introduction of a second negative charge would result in a less favorable free energy change. Instead of reporting a linear dependence, the $\Delta G_{protein}(pH)$ curves each have an inflection point and two asymptotic values, corresponding to the protonation free energy of one residue, while the other residues remain in the same protonation state (Figure 7b, top). These asymptotic values correspond to the microscopic pK_a values (Figure 7b, middle), which put a bound on the range of computed pK_a values.

At low pH, both A_1 and B_2 are protonated, with large, unfavorable ΔG values (Figure 7b, top). As the pH increases, ΔG of deprotonating B_2 becomes more favorable, reaching $\Delta G = 0$ near pH $\approx$ 3; at this point, B_2 begins to deprotonate. This deprotonation will result in a more unfavorable protonation free energy for A_1, as evidenced by the flattening in the ΔG curve and an increase in the apparent pK_a of A_1: the formation of B_2^- makes the formation of A_1^- less favorable. Because the reference pK_a° values of A_1 and B_2 are similar, this flattening occurs almost

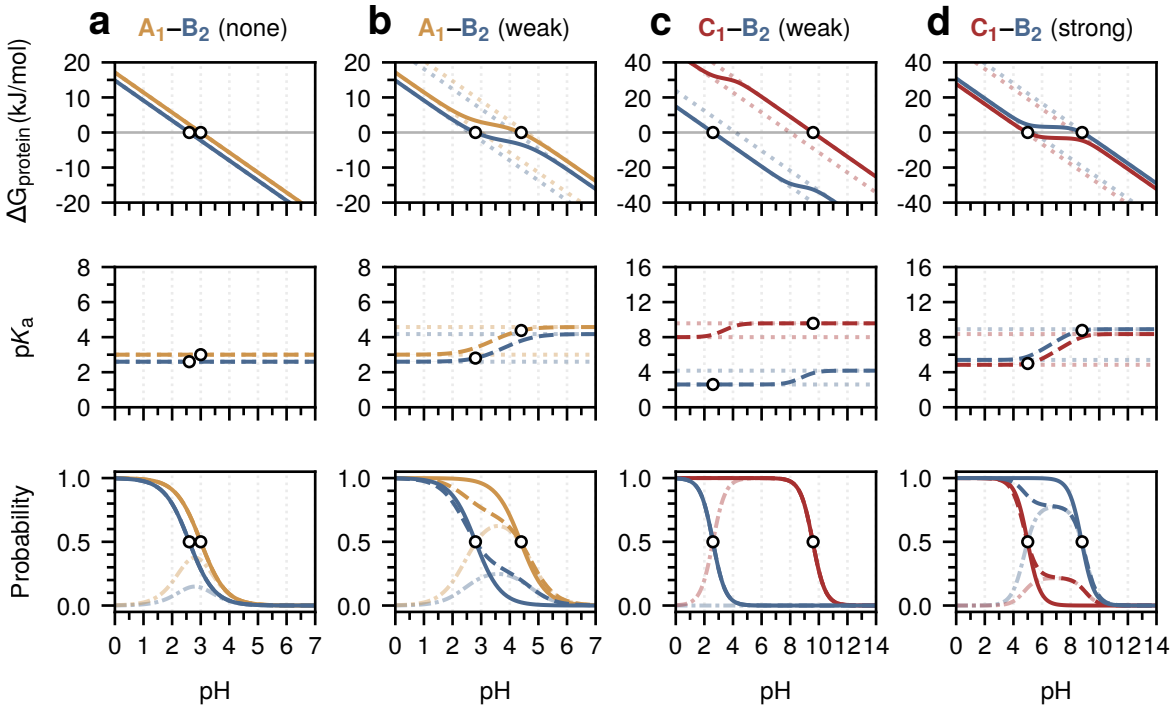

Figure 7: pH-dependent free energy, pK_a, and protonation probability curves: toy systems. The upper plots (dashed) depict $\Delta G_{protein}$(pH) (Equation 15), which varies as a function of pH between its asymptotic values (dotted). The zero point of this curve (single dot) is used to resolve the corresponding pK_a value. The middle plots (dashed) depict the pH-dependent pK_a value (Equation 16). Each residue has two limiting pK_a values (dotted) which correspond to the cases when the other coupled residue is protonated or deprotonated. As the pH changes, the probability that the other residue is deprotonated shifts, resulting in a pH-dependent pK_a. The lower plots depict several probabilities. The dashed lines correspond to the protonation probabilities of the individual sites; these are a composite probability of the doubly protonated (i.e., $\langle A^H B^H \rangle$) and singly protonated (i.e., $\langle A^H B^- \rangle$) microstates (see Equation 18). Singly protonated probabilities are indicated with dashed-dotted lines. The solid lines correspond to the standard Henderson-Hasselbalch curve computed for the pK_a determined by the zero point of $\Delta G_{protein}$. Because we resolve the pK_a at a single pH, these curves are sigmoidal. Observe that with no coupling (left column), the pK_a values are constant; however, when coupling is introduced, this is no longer the case. Columns **a**–**d** correspond to four different coupling scenarios.

simultaneously with that of B_2. In this regime, the $\Delta G_{protein}$ for both residues changes slower and, in this example, remains relatively close to zero. We can think of this as the pH range over which the groups buffer each other, altering the favourability of protonation. As the pH continues to increase, a linear dependence is restored. Construction of the titration curves computed from the microscopic pK_a values reveals the coupling between residues (Figure 7b, bottom). The non-sigmoidal form of the curves follows from the fact that the singly protonated microstates for both residues occur with a similar probability.

Model 3: Different reference pK_a° values: weak coupling

Consider the system given in Figure 8. In this example, the $\Delta\Delta G$ values along the branches are the same as in Example 2; however, the reference pK_a° values have changed. We now consider the coupling between a residue C that has a reference value 5 pK units higher than B. Although $\Delta G_{protein}$(pH) does report a non-linear dependence, the large difference in reference values means that the pH effects dominate; both inflection points of the free energy curves and the inflection point of the pH-dependent pK_a values

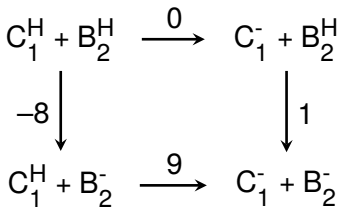

Figure 8: Thermodynamic cycle for Toy Model 3. ΔΔG values (kJ/mol) are indicated along the branches.

occur far from one another. As a result, the titration curves computed from the microscopic pK_a values suggest that there is no coupling between residues. The singly protonated microstates for each residue occur with a dramatically different probability, that is, residue B_2 is never protonated while C_1 deprotonates.

Model 4: Different reference pK_a° values: strong coupling

Consider the system given in Figure 9. Relative to

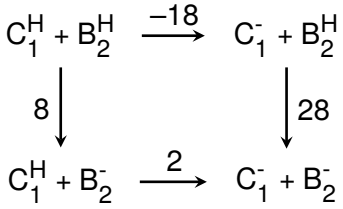

Figure 9: Thermodynamic cycle for Toy Model 4. ΔΔG values (kJ/mol) are indicated along the branches.

Example 3 we have altered the ΔΔG values along the branches, while maintaining the reference pK_a° values. Here, the interaction between the residues is much stronger and a more favorable ΔΔG is assigned to the initial deprotonation of C_1. Unlike in Example 3, where the titration events occurred far from one another, both the ΔG and pK_a curves resemble those in Example 2. However, unlike Example 2 the reference pK_a° values differ by 5 pK units; in this case, the reference pK_a gap is compensated by significant shifts in the ΔΔG values. Here, also note that the buffer region over which coupling occurs is larger than in Example 2 (Figure 7b, middle). In this region, the effect of pH on ΔG is much less pronounced, and the ΔG of protonation remains relatively constant (i.e., slope of $\Delta G_{protein}$ is zero), within 1 pK unit of zero.

As in Example 2, constructing the titration curves computed from the microscopic pK_a values reveals the coupling between residues (Figure 7c, bottom). Here, plateaus are evident for both residues over pH $\in [6, 9]$, implying the existence of both singly protonated microstates. Moreover, the residue with the higher reference pK_a°, perhaps counterintuitively, titrates first.

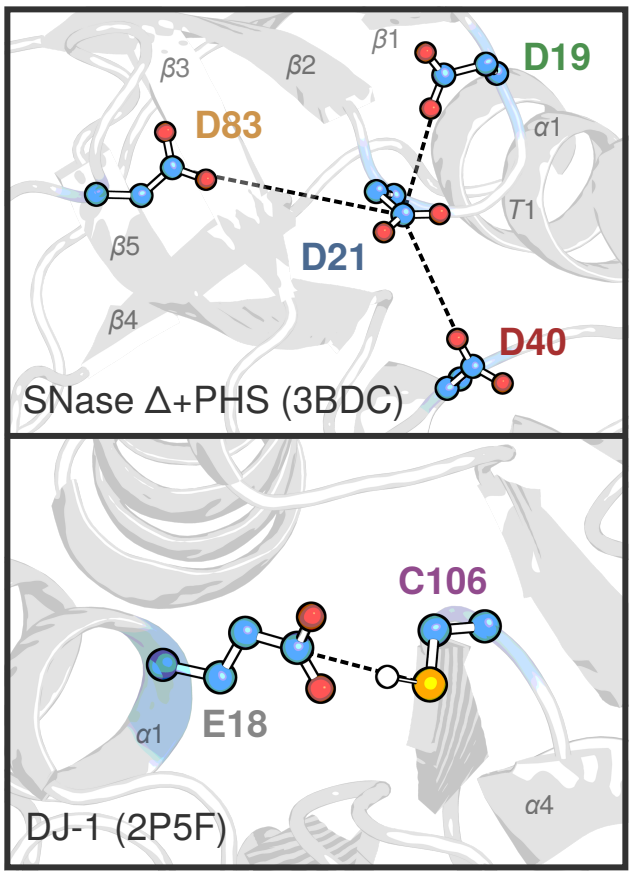

Figure 10: Coupled residues considered. Upper: Δ + PHS SNase (PDB: 3BDC); lower: monomeric DJ-1 (PDB: 2P5F). Carbon atoms are shown in blue, oxygen atoms in red, and sulfur atoms in yellow. Residue dyads are considered independently and are indicated with dashed lines.

Application to proteins

Non-equilibrium alchemy

We consider four potential dyads: three from the SNase variant[30] (PDB: 3BDC[30]) and one from protein deglycase DJ-1[31] (PDB: 1P5F[32]). These pairs are D19–D21, D21–D40, and D21–D83 in

SNase and E18–C106 in DJ-1. In SNase, D19 and D21 are spatially adjacent to each other within the turn ($T1$) between $\beta1$ and $\beta2$, while D40 is located in the turn between $\beta3$ and $\alpha1$, which puts it close to D21 but farther away than D19 (Figure 10, top). D83 is located in the long linker between $\beta4$ and $\beta5$, and is the farthest from D21 of the dyads considered here. Although sequentially distant, E18 and C106 in DJ-1, located within $\alpha1$ and near $\alpha4$, respectively, are directly adjacent in space (Figure 10, bottom).

The three SNase pairs exhibited different levels of coupling: the D21–D83 dyad showed almost no coupling (Figure 11a), as evidenced by the absence of inflection points in the ΔG and pK_a curves, the D21–D40 showed moderate coupling (Figure 11b), and the D21–D19 dyad showed significant coupling (Figure 11c). In this latter case, as in the second toy example, the residues clearly acted to buffer one another, resulting in a flattening of ΔG for both curves around pH $\in [3, 5]$. Moreover, computing protonation probabilities for the individual sites revealed a non-sigmoidal form of the curves. We note that the coupling of these residues was also observed in the experiment. The calculated pK_a values for SNase were: D19: 3.69 $\pm$ 0.09, D21: 5.34 $\pm$ 0.16, D40: 3.88 $\pm$ 0.05, D83: 1.17 $\pm$ 0.16, in good agreement with experiment (D19: 2.21 $\pm$ 0.01, D21: 6.54 $\pm$ 0.02, D40: 3.87 $\pm$ 0.09, and D83: < 2.2) (Figure 12a).

In the case of E18–C106, again a flattening in the ΔG buffer region and, evidently, non-sigmoidal individual site pK titration curves suggested a coupling between the residues (Figure 11d). Here, computing the adjusted pK_a downshifts C106 from 12.18 $\pm$ 0.07 to 8.84 $\pm$ 0.04, bringing the estimate closer to the experimental value of 5.4 $\pm$ 0.2 (Figure 12a). However, this still leaves more than a 3 pK unit discrepancy between calculation and experiment. Previous work on homodimeric DJ-1 has revealed that two arginine residues (R48 and R28 from the other monomer) facilitate anion binding, which results in pK_a elevation[31] (Figure S1, right). In our simulations only positive counterions (i.e., no salt concentration) were present; however, through-space interactions as a result of a second arginine may also play a role in affecting the pK_a. To this end, we probed the pK_a of monomeric DJ-1. We found similar qualitative agreement in the curves between the dimeric and monomeric forms; however, rather than raising the pK_a, the elimination of the second arginine shifted the pK_a of C106 down to 6.78 $\pm$ 0.19 (Figure S1, left).

We note that in both cases an exceptionally high pK_a for E18 is predicted. While previous work on DJ-1 has suggested that E18 is protonated over the titration regime of C103[31] and glutamate residues have been reported with pK_a values greater than 9,[41] it would seem improbable that the pK_a is actually this high. Structurally, a second glutamate (E18) and a nearby histidine (H126) likely play roles within the lower pH regime (i.e., $<$ 7). We preface the following section by noting that a high glutamate pK_a value is also suggested by the CpHMD simulations and PropKa as described further.

Constant-pH molecular dynamics and PropKa

To further investigate the reliability of our approach, we performed constant-pH molecular dynamics simulations of the same systems. Here, we found good agreement with the alchemical calculations. In the case of D21–D83, a single inflection point was implied by the model selection (Equation 21), whereas for D21–D40 and D19–D21, a double fit, rather than a single fit to the Henderson-Hasselbalch curve (Equation 22) was implied (Figure S2). Resolving the pK_a values implied by the fits, we computed: 0.62 $\pm$ 0.10 pK for D83, 3.08 $\pm$ 0.07/4.48 $\pm$ 0.06 pK for D40, and 2.47 $\pm$ 0.08/4.94 $\pm$ 0.16 pK for D19. In each case, different values for D21 were calculated: 5.44 $\pm$ 0.15 (coupled with D83), 3.86 $\pm$ 0.11/7.96 $\pm$ 0.81 (coupled with D40), and 3.25 $\pm$ 0.09/6.24 $\pm$ 0.26 (coupled with D19). Assigning the double fit pK_a values to the individual residues and averaging over D21, we have a tentative assignment: D19: 2.86 $\pm$ 0.09 pK, D21: 6.55 $\pm$ 0.41 pK, D40: 3.47 $\pm$ 0.07 pK, D83: 0.62 $\pm$ 0.10 pK, which was in very good agreement with experiment (Figure 12a).

Regarding DJ-1, the large pK_a value of E18 implied by free energy calculations is also suggested by the constant-pH simulations on monomeric DJ-1, where values of 9.48 $\pm$ 0.42 pK and 20.99 $\pm$ 0.05 pK were calculated in the cases

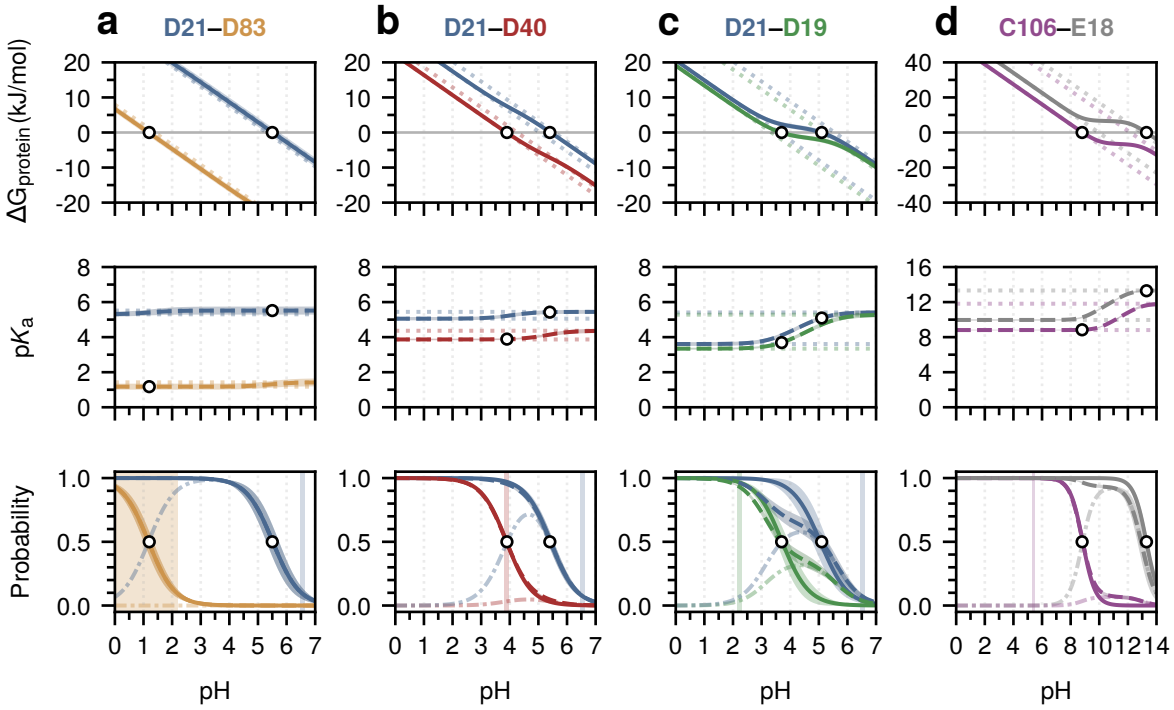

Figure 11: pH-dependent free energy, pK_a, and protonation probability curves: protein systems. The upper plots (dashed) depict $\Delta G_{protein}$(pH) (Equation 15), which varies as a function of pH between its asymptotic values (dotted). The zero point of this curve (single dot) is used to resolve the corresponding pK_a value. The middle plots (dashed) depict the pH-dependent pK_a value (Equations 16). Each residue has two limiting pK_a values (dotted) which correspond to the cases when the other coupled residue is protonated or deprotonated. As the pH changes, the probability that the other residue is deprotonated shifts, resulting in a pH-dependent pK_a. The lower plots depict several probabilities. Dashed lines correspond to the protonation probabilities of the individual sites; these are a composite probability of the doubly protonated (i.e., $\langle A^H B^H \rangle$) and singly protonated (i.e., $\langle A^H B^- \rangle$) microstates (see Equation 18). The singly protonated probabilities are indicated with dashed-dotted lines. Solid lines correspond to the standard Henderson-Hasselbalch curve computed for the pK_a determined by the zero point of $\Delta G_{protein}$. Because we resolve the pK_a at a single pH, these curves are sigmoidal. Observe that with no coupling (left column), the pK_a values are constant; however, when coupling is introduced, this is no longer the case. Vertical spans in the lower row indicate experimental pK_a values and uncertainties (note that D83 has a pK_a < 2.2). Error bands were bootstrapped. **a**–**c** correspond to the SNase + ΔPHS system, while **d** corresponds to the DJ-1 system.

of protonated C106 and deprotonated C106, respectively. Here, only a single fit to Equation 21 was performed and the Hill coefficient was not fixed to n = 1. C106 was not probed as cysteine residues are not yet supported by the current CpHMD implementation.[40]

We close this section by briefly comparing our results with the popular computational pK_a predictor PropKa.[42] Overall, for PropKa, we find agreement for the SNase dyads, but a worse accuracy for C106 in DJ-1. Specifically, estimates from PropK_a were: D19: 3.21, D21: 5.44, D40: 4.30, D83: 1.21, and C106: 14.19 (Figure 12a). In the case of E18 in DJ-1, PropKa estimates a pK_a of 8.73 and 7.38 for the homodimeric and monomeric forms, respectively.

Here, our NEQ approach provided estimates of both D40 and C106 that were in closer agreement with the experimental values while exhibiting comparable performance on the other three residues. Considering the overall performance we found that the introduction of coupling dramatically

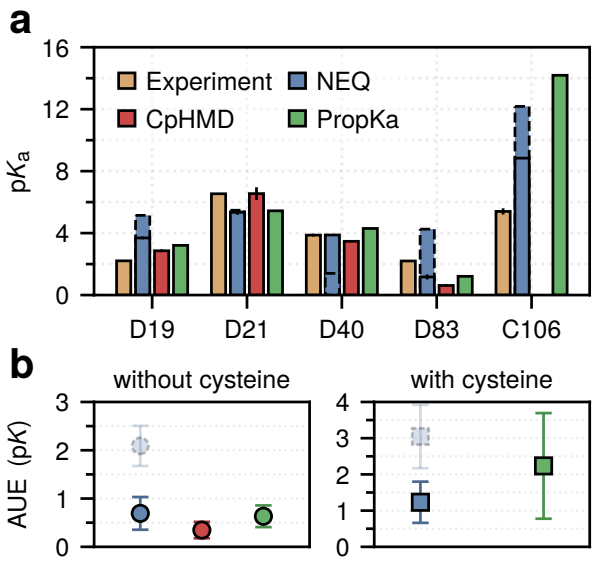

Figure 12: Performance of various methods for calculating protein pK_a values. Three methods are compared with experiment: NEQ with (solid) and without (dashed/transparent) coupling accounted for; constant-pH MD (CpHMD); and PropKa. Note that the pK_a of D83 is < 2.2 pK. **a.** Residue-wise performance **b.** Overall performance with or without cysteine included. Bootstrapped standard errors are depicted.

improves agreement with experiment reducing our NEQ average unsigned error from 2.10 ± 0.42 to 0.69 ± 0.34 when cysteine is excluded and from 3.05 ± 0.88 to 1.23 ± 0.57 when it is included; with regard to the former, this performance was comparable to PropKa (0.63 ± 0.22) (Figure 12b). We also found that CpHMD could accurately resolve the four aspartates within 0.5 pK. It is probable that both a limited training set and a less frequent dyad (i.e., a large difference in reference pK_a° values) results in the markedly poorer estimate for C106 from PropKa.

Discussion

The importance of accounting for residue coupling is multifaceted and particularly relevant in the context of enzymatic active sites that are often enriched in protonatable residues. The theoretical formalism to describe such couplings in polyprotic acids has been detailed by Ullman.[18] In his work, Ullman follows an approach of defining equilibrium protonation constants for all microstates and subsequently derives partition functions fully describing the thermodynamics of these systems.

However, in our work, we follow a different path and describe the titration of coupled sites starting from the double free energy difference of protonation in the protein with respect to a reference state in water. This approach is particularly relevant in the context of alchemical free energy calculations, which give access only to such $\Delta\Delta G$ values.

Here, we explore both toy examples and real protein systems where the buffering of a residue dyad maintained the free energy of protonation near zero (e.g., DJ-1: C106-E18). In various protein contexts, tuning the local residue environment surrounding pairs or groups of titratable residues to create large buffer regions over which the ΔG of protonation is close to zero could make the binding of a substrate more than sufficient to significantly alter the (de)protonation of a residue and ultimately the enzymatic activity.

Comparison with a recent GROMACS-based CpHMD implementation[40] revealed a good pK_a prediction accuracy for coupled residues. This result suggests that both CpHMD and alchemical free energy methods can resolve pK_a values in both coupled and uncoupled contexts. A second comparison with PropKa suggested that although accurate estimates could be made for SNase, the pK_a of C106 in DJ-1 was significantly overestimated, possibly due to the limited cysteine data in the PropKa training set and the implicit assumption that E18 is deprotonated at the pH where C106 deprotonates.

The role of MD simulations and free energy calculation methods such as those employed here may provide insight not readily accessible to conventional prediction methods. One particular insight, namely the pK_a values of coupled residues, requires careful consideration of the role of the protonatable residues nearby. Moreover, given that these residues are often found at the active site — frequently the target of engineered therapeutics — the relevance of this problem extends beyond basic research.

The ability to seamlessly and consistently integrate such pH-dependent calculations into existing alchemical free energy workflows may prove invaluable for accurately resolving

binding affinities or enzyme activities and thermostabilities.

To this end, we have elsewhere investigated the ability of NEQ free energy calculations to compute a large number of pK_a values in a variety of protein contexts. As with the results here, our approach showed strong performance, further suggesting the potential for a consistent integration of pH-dependent calculations into a broader NEQ free energy framework (*Manuscript in preparation*).

Acknowledgement C.J.W. thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Government of Ontario for funding.

Supporting Information

The Supporting Information is available free of charge at [link].

References

(1) Talley, K.; Alexov, E. On the pH-optimum of activity and stability of proteins. *Proteins* **2010**, *78*, 2699–2706.

(2) Srivastava, J.; Barreiro, G.; Groscurth, S.; Gingras, A. R.; Goult, B. T.; Critchley, D. R.; Kelly, M. J. S.; Jacobson, M. P.; Barber, D. L. Structural model and functional significance of pH-dependent talin–actin binding for focal adhesion remodeling. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14436–14441.

(3) Boyken, S. E.; Benhaim, M. A.; Busch, F.; Jia, M.; Bick, M. J.; Choi, H.; Klima, J. C.; Chen, Z.; Walkey, C.; Mileant, A.; Sahasrabuddhe, A.; Wei, K. Y.; Hodge, E. A.; Byron, S.; Quijano-Rubio, A.; Sankaran, B.; King, N. P.; Lippincott-Schwartz, J.; Wysocki, V. H.; Lee, K. K.; Baker, D. De novo design of tunable, pH-driven conformational changes. *Science* **2019**, *364*, 658–664.

(4) Ripstein, Z. A.; Vahidi, S.; Rubinstein, J. L.; Kay, L. E. A pH-Dependent Conformational Switch Controls *N. meningitidis* ClpP Protease Function. *J. Am. Chem. Soc.* **2020**, *142*, 20519–20523.

(5) Kaptan, S.; Assentoft, M.; Schneider, H. P.; Fenton, R. A.; Deitmer, J. W.; MacAulay, N.; de Groot, B. L. H95 Is a pH-Dependent Gate in Aquaporin 4. *Structure* **2015**, *23*, 2309–2318.

(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) 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.

(8) 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.

(9) Yao, X.; Chen, C.; Wang, Y.; Dong, S.; Liu, Y.-J.; Li, Y.; Cui, Z.; Gong, W.; Perrett, S.; Yao, L.; Lamed, R.; Bayer, E. A.; Cui, Q.; Feng, Y. Discovery and mechanism of a pH-dependent dual-binding-site switch in the interaction of a pair of protein modules. *Sci. Adv.* **2020**, *6*, eabd7182.

(10) Warshel, A. Calculations of enzymic reactions: calculations of pKa, proton transfer reactions, and general acid catalysis reactions in enzymes. *Biochemistry* **1981**, *20*, 3167–3177.

(11) 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* **2010**, *79*, 685–702.

(12) 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.

(13) Dodson, G. Catalytic triads and their relatives. *Trends Biochem. Sci.* **1998**, *23*, 347–352.

(14) Du, Z.; Zheng, Y.; Patterson, M.; Liu, Y.; Wang, C. pKa Coupling at the Intein Active Site: Implications for the Coordination Mechanism of Protein Splicing with a Conserved Aspartate. *J. Am. Chem. Soc.* **2011**, *133*, 10275–10282.

(15) 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.

(16) Edsall, J. T.; Wyman, J. *Biophysical Chemistry: Thermodynamics, Electrostatics, and the Biological Significance of the Properties of Matter*; Academic Press Inc., 1958.

(17) Onufriev, A.; Case, D. A.; Ullmann, G. M. A Novel View of pH Titration in Biomolecules. *Biochemistry* **2001**, *40*, 3413–3419.

(18) Ullmann, G. M. Relations between Protonation Constants and Titration Curves in Polyprotic Acids: A Critical View. *J. Phys. Chem. B* **2003**, *107*, 1263–1271.

(19) Tanford, C.; Kirkwood, J. G. Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.

(20) Sharp, K. A.; Honig, B. Electrostatic Interactions in Macromolecules: Theory and Applications. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301–332.

(21) Gapsys, V.; Michielssens, S.; Peters, J. H.; Groot, B. L. d.; Leonov, H. *Molecular Modeling of Proteins*; Springer, 2015; pp 173–209.

(22) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245–268.

(23) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Phys. Rev. Lett.* **2003**, *91*, 140601.

(24) Crooks, G. E. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E* **1999**, *60*, 2721–2726.

(25) 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.* **2016**, *55*, 7364–7368.

(26) 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.

(27) 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*, 1–13.

(28) Hill, T. L. *Cooperativity Theory in Biochemistry: Steady-State and Equilibrium Systems*, 1st ed.; Springer Series in Molecular and Cell Biology; Springer: New York, NY, 1985.

(29) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated protein structure and topology generation for alchemical perturbations. *J. Comput. Chem.* **2014**, *36*, 348–354.

(30) 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* **2009**, *77*, 570–588.

(31) Witt, A. C.; Lakshminarasimhan, M.; Remington, B. C.; Hasim, S.; Pozharski, E.; Wilson, M. A. Cysteine pKa Depression by a Protonated Glutamic Acid in Human DJ-1. *Biochemistry* **2008**, *47*, 7430–7440.

(32) Wilson, M. A.; Collins, J. L.; Hod, Y.; Ringe, D.; Petsko, G. A. The 1.1-Å resolution crystal structure of DJ-1, the protein mutated in autosomal recessive early onset Parkinson's disease. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9256–9261.

(33) 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* **2016**, *14*, 71–73.

(34) 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.

(35) Gunsteren, W. F. V.; Berendsen, H. J. C. A Leap-frog Algorithm for Stochastic Dynamics. *Mol. Sim.* **1988**, *1*, 173–185.

(36) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(37) 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.

(38) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2007**, *4*, 116–122.

(39) Gapsys, V.; de Groot, B. L. On the importance of statistics in molecular simulations for thermodynamics, kinetics and simulation box size. *eLife* **2020**, *9*.

(40) Aho, N.; Buslaev, P.; Jansen, A.; Bauer, P.; Groenhof, G.; Hess, B. Scalable Constant pH Molecular Dynamics in GROMACS. *J. Chem. Theory Comput.* **2022**,

(41) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A summary of the measured pKa values of the ionizable groups in folded proteins. *Prot. Sci.* **2008**, *18*, 247–251.

(42) 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.

Supporting Information

Carter J. Wilson,[1,2] Bert L. de Groot,[3] and Vytautas Gapsys[3,4,*]

[1]Department of Mathematics, The University of Western Ontario, N6A 5B7, London, Canada
[2]Centre for Advanced Materials and Biomaterials Research (CAMBR),
The University of Western Ontario, N6A 5B7, London, Canada
[3]Computational Biomolecular Dynamics Group, Department of Theoretical and Computational Biophysics, Max Planck Institute for Multidisciplinary Sciences, 37077 Göttingen, Germany
[4]Computational Chemistry, Janssen Research & Development, Janssen Pharmaceutica N. V., Turnhoutseweg 30, B-2340 Beerse, Belgium

*vgapsys@gwdg.de

Supplemental Figures

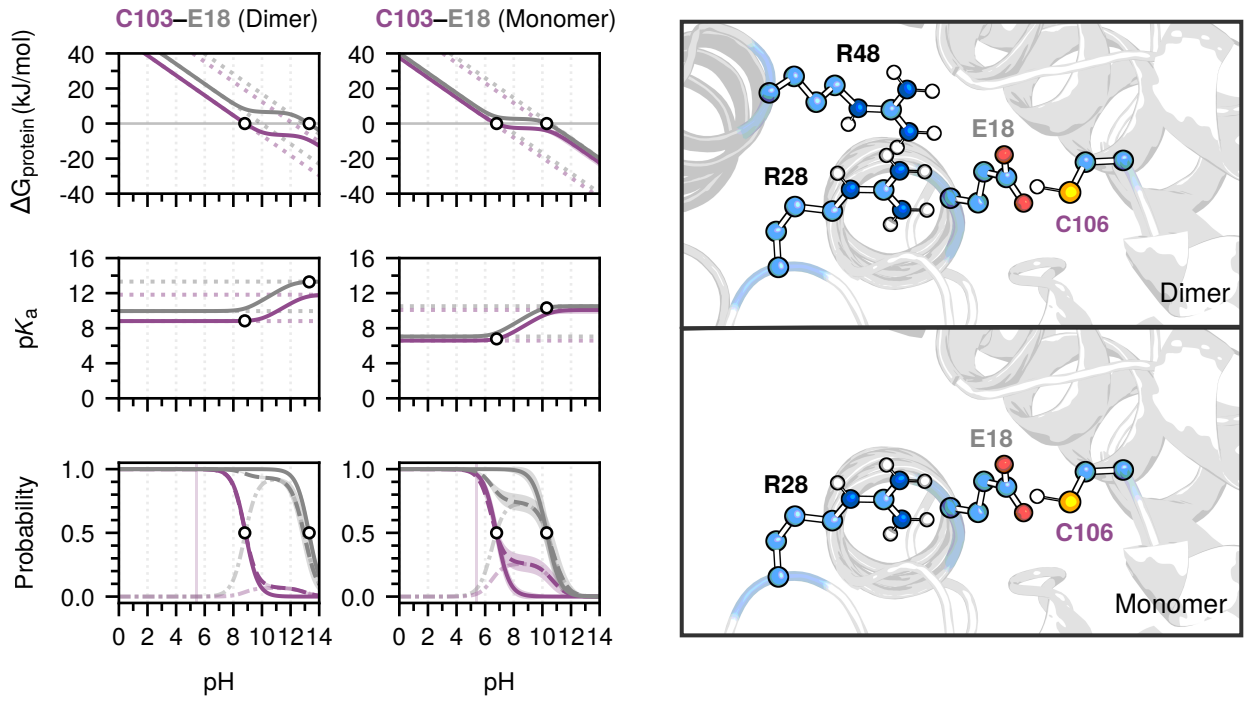

Figure S1: Homodimeric and monomeric structures of DJ-1 (PDB: 2P5F) are depicted alongside the computed pK_a values for C106 and E18. The addition of a second nearby arginine (R48) in the homodimer significantly shifts the pK_a values. pH-dependent pK_a curves (middle row) and the corresponding pK_a values as determined from the zero point of the $\Delta G_{protein}$(pH) curves (upper row). Dotted lines (upper/middle rows) correspond to the the microscopic values. Protonation probability curves (lower row) based on the microstate probability equations (dashed), pK_{int} (dotted), and pK_a (solid). Vertical spans in the lower row indicate experimental pK_a values and uncertainties. Carbon atoms are shown in blue, oxygen atoms in red, and sulfur atoms in yellow. Error bands were bootstrapped.

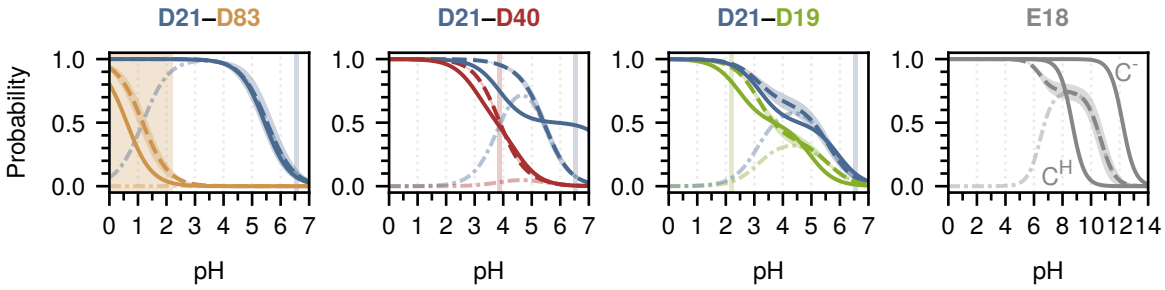

Figure S2: Solid lines correspond to the CpHMD computed pK_a protonation probabilties based on Equations 21 and 22 in the main text. Dashed lines correspond to the protonation probabilities of the individual sites; these are a composite probability of the doubly protonated (i.e., $\langle A^H B^H \rangle$) and singly protonated (i.e., $\langle A^H B^- \rangle$) microstates (see Equation 18). The singly protonated probabilities are indicated with dashed-dotted lines. Vertical spans indicate experimental pK_a values and uncertainties (note that D83 has a p$K_a < 2.2$). Error bands were bootstrapped. In the case of E18, CpHMD simulations were run with C106 protonated (C^H) or deprotonated (C^-).

TOC Graphic