Journal of Computer-Aided Molecular Design (2020) 34:561–573
https://doi.org/10.1007/s10822-020-00280-7

Standard state free energies, not pK_as, are ideal for describing small molecule protonation and tautomeric states

M. R. Gunner[1] · Taichi Murakami[1] · Ariën S. Rustenburg[2,3,5] · Mehtap Işık[2,4] · John D. Chodera[2]

Received: 21 October 2019 / Accepted: 8 January 2020 / Published online: 12 February 2020

Abstract

The pK_a is the standard measure used to describe the aqueous proton affinity of a compound, indicating the proton concentration (pH) at which two protonation states (e.g. A^- and AH) have equal free energy. However, compounds can have additional protonation states (e.g. AH_2^+), and may assume multiple tautomeric forms, with the protons in different positions (microstates). Macroscopic pK_as give the pH where the molecule changes its total number of protons, while microscopic pK_as identify the tautomeric states involved. As tautomers have the same number of protons, the free energy difference between them and their relative probability is pH independent so there is no pK_a connecting them. The question arises: What is the best way to describe protonation equilibria of a complex molecule in any pH range? Knowing the number of protons and the relative free energy of all microstates at a single pH, ΔG°, provides all the information needed to determine the free energy, and thus the probability of each microstate at each pH. Microstate probabilities as a function of pH generate titration curves that highlight the low energy, observable microstates, which can then be compared with experiment. A network description connecting microstates as nodes makes it straightforward to test thermodynamic consistency of microstate free energies. The utility of this analysis is illustrated by a description of one molecule from the SAMPL6 Blind pK_a Prediction Challenge. Analysis of microstate ΔG°s also makes a more compact way to archive and compare the pH dependent behavior of compounds with multiple protonatable sites.

Keywords SAMPL6 · pK_a · Tautomer · Protonation state · pH titration · Multiprotic

Introduction

Acids and bases in solution bind and release protons in a process that changes their molecular charge distribution, influencing solubility, and molecular recognition and reactivity. Many biological and bio-active molecules have pK_as in the physiological pH range, existing in a mixture of protonation and tautomeric states. The knowledge of which protonation states are energetically accessible is important for the design of molecules with desired function [1, 2]. The binding or loss of a proton represents one of the simplest reactions, so the calculation of the relative free energy of protonation states as a function of pH represents a powerful test of biomolecular modeling methodologies [3, 4].

The recent SAMPL6 Blind pK_a Prediction Challenge focused on the prediction of pK_as for 24 small molecules [5]. As the participant submissions were analyzed and compared it became apparent that the description of the free energy landscape of molecules with multiple protonation states is not simple and that lists of many pK_a values is in fact not the

✉ M. R. Gunner
mguner@ccny.cuny.edu

[1] Department of Physics City College of New York, New York, NY 10031, USA

[2] Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[3] Graduate Program in Physiology, Biophysics and Systems Biology, Weill Cornell Medical College, New York, NY 10065, USA

[4] Tri-Institutional PhD Program in Chemical Biology, Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, NY 10065, USA

[5] Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY 10065, USA

best way to describe the behavior of the molecule as a function of pH. Attempting to capture all necessary information regarding pK_a predictions, the SAMPL6 pK_a Challenge supported three different reporting schemes (submission types): microscopic pK_a values (type I), fractional populations of microstates with respect to pH (type II), and macroscopic pK_as (type III). These reporting schemes captured different aspects of the predictions, however none of them has proved to be optimal. Here we present a different reporting scheme that provides a complete and concise description of the thermodynamic behavior of molecules with multiple protonation and tautomeric states, and allows the derivation of pK_as.

Proteins can have innumerable protonation and tautomeric microstates

The complexity of protonation equilibria of a molecule can vary enormously depending on the number of possible titratable groups and the interactions amongst them. Thus, a simple molecule with a single titratable group has a single well-defined pK_a, which is the pH where the species with different numbers of protons have the same free energy and thus the same concentration. For a single protonatable group the reaction is:

$$AH = A^- + H^+ \tag{1a}$$

Defining the pK_a as $-\log_{10} K_{eq}$ leads to:

$$pK_a = pH - \log_{10} \frac{A^-}{AH} \tag{1b}$$

We will define a protonation macrostate by the total charge of the molecule while the microstate defines the specific protonation and tautomeric state of all protonatable sites. As the number of protonatable sites in a molecule increases, the number of possible microstates the molecule can access increases. Proteins and other (bio)polymer polyelectrolytes [6] can have many acidic and basic substituents. If we only consider protonatable sites that can gain or lose a single proton (A^- vs. AH or B vs. BH^+) then there are 2^n different distributions of protonation states for n protonatable groups. On average, 25% of protein residues are either Asp, Glu, Lys, or Arg [7], providing a very large number of possible microstates. Long-range electrostatic interactions lead to the ionization of all residues being interdependent [8]. Computational tools have been developed that view the protein environment as perturbing the pK_as individual residues would have in solution [9–11]. Given the huge number of possible microstates, Metropolis Monte Carlo sampling is typically used to sample the Boltzmann distribution of protonation states at each pH [11]. Thus, proteins have many protonatable sites, where the interactions are not negligible, but can be treated as a sum of separable, individual interactions. Typically, calculations consider a relatively rigid protein and calculate the electrostatic interactions with the Poisson-Boltzmann equation of continuum electrostatics. Evolving methods allow the protein to move via classical MD simulations, with the protonation change sampled either by a separate MC analysis or by lambda dynamics within the MD trajectory [12–15]. However, these methods will not work to define the pH dependent behavior of small molecules with multiple protonation states.

SAMPL6 pK_a challenge targets are small molecules with multiple protonation and tautomeric microstates

Organic molecules with multiple protonatable sites play roles in metabolism and are design targets for drugs [1, 2]. The pH dependent behavior of these molecules is at a level of complexity between a single protonatable group (Eq. 1) and a complex polyelectrolyte such as a protein. The SAMPL6 pK_a Challenge chose 24 extended organic compounds as a test case for a blind prediction of molecule pK_as [5, 16]. (https://github.com/samplchallenges/SAMPL6/tree/master/physical_properties/pKa/microstates). All SAMPL6 molecules have multiple potential protonation macrostates (states with different total charge) as well as different energetically accessible tautomer states (same total charge but different protonation site), each of which represents a defined microstate. There are three to six protonatable sites in each molecule, substantially fewer than in proteins, but still sufficient to generate tens of protonation and tautomeric microstates (Fig. 1) [4, 5, 17]. However, the coupling amongst the protonatable sites in these molecules does not allow separation into independent units, with simply summable interactions. Rather, each molecule must be treated as a whole with its microstate energy a function of the proton distribution and molecular conformation. The prediction methods used in submissions to the SAMPL6 pK_a Challenge range from knowledge-based empirical methods to detailed quantum mechanical simulations. Several recent papers have described some of the results [16, 18–23] and reference [17] provides an analysis of the predictions submitted for all molecules.

Figure 1 shows the network of 11 considered protonation and tautomeric microstates of the SAMPL6 pK_a Challenge molecule SM07, which will be described in detail here. While it is theoretically possible to enumerate additional microstates, this set combines microstates suggested by SAMPL6 pKa Challenge organizers (using Epik from the Schrodinger Suite v3.4 and QUACPAC from the OpenEye Toolkit v2017.Feb.1 plus any additional microstates included in SAMPL6 challenge submissions). The protons

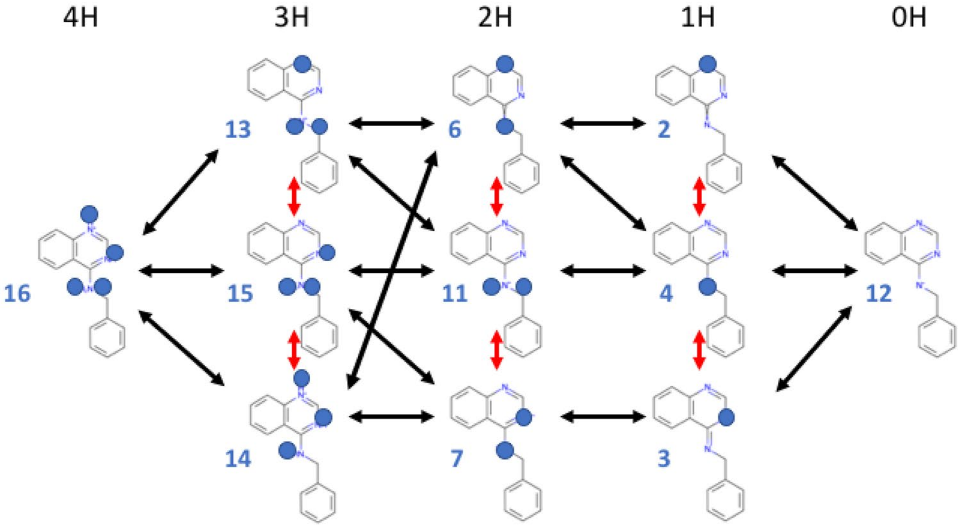

Fig. 1 Network of protonation and tautomeric microstates for SAMPL6 pK_a challenge target SM07

of interest are denoted by blue balls on top of the nitrogen which is the proton acceptor. Each column has a different number of dissociable protons, indicated at the top. All microstates in a column are tautomers with the same total charge; their vertical order is arbitrary. All the tautomers in a column contribute to the macrostates with 4 (4H) to zero (0H) dissociable protons. Black double-headed arrows indicate pK_as that were reported in the SAMPL6 Challenge; red arrows are the transitions between tautomers. Some transitions, such as between 13 and 7 required tautomer changes. Likewise, transitions between tautomers in the top and bottom rows (e.g. between microstates 2 and 3) are also well-defined, but are not shown here for clarity. The numbers associated with each microstate simplify the microstate IDs assigned by the SAMPL6 pKa Challenge, which have the form: SM07_microXXX, where XXX are three digits. For example, microstate 4 in this figure corresponds to SM07_micro004.

Figure 1 shows many closed reaction cycles. One example, starts with state 4 (1H); the shift of a proton position leads to tautomeric microstate 2 (1H); the loss of a proton generates microstate 12 (0H); and proton binding regenerates microstate 4. In a network that is thermodynamically consistent the summed change in free energy for these three reactions should equal zero. The network shows cycles, including many with 3 microstates as well as larger ones such as those that connect microstates 12 (0H) and 16 (4H) through different tautomers of the intermediate protonation states. Thermodynamic consistency can provide a test of a set of calculations for a given molecule.

What information was requested for the SAMPL6 pK_a challenge?

Three types of submissions for predictions were requested: microscopic pK_as for related microstate pairs (type I), fractional microstate populations in the pH interval 2 to 12 (type II), and macroscopic pK_as (type III). We will see that type I and type II are formally identical as long as the same microstates are included in both descriptions and that type III misses important information needed to see if a numerical pK_a that matches an experimental value captures the correct protonation and tautomeric states.

Macroscopic pK_a entries were reported (type III submissions). Macroscopic states combine all the (tautomer) microstates that have the same number of protons and thus the same net charge. Macroscopic pK_as are closest to experiments that monitor the proton uptake of the molecule as a function of pH, such as electrochemical titrations or spectrophotometric titrations [5]. However, the macroscopic predictions did not require an assignment of which microstates are involved or even how many protons are associated with the beginning and end state. Thus, it may not be clear if the transition is, for example, between A^- and AH or between AH and $AH_2^{+\cdot}$ Unfortunately, this ambiguity is also a problem for many of the experimental measurements of these complex molecules. Thus, a spectroscopic or potentiometric titration provides information on the pK_a value of a proton binding event without knowing the macrostate (charge state) or microstate(s) (tautomer(s) with that charge) that are connected by the pH-dependent transition. This lack of specificity made it difficult to determine if different methods predicting a similar pK_a were referring to transitions between even the same macrostates of the molecule [5].

The minimum information needed to describe the network of protonation and tautomeric states: microstate ΔG°s at one pH provides one number to rule them all

The complexity of the SAMPL6 protonation and tautomer microstates for each molecule led to an unanticipated open question: What is the best way to report the predictions? The ideal description should provide the minimum information needed so the free energy landscape for all the protonation and tautomeric microstates in a network, such as that shown in Fig. 1, can be described at each pH. It should allow easy comparison with experimental measurements and between multiple predictions for the same system. It should define the distribution of tautomeric microstates for each protonation macrostate and include information about high energy microstates, which could become important in a specific binding pocket or in a reaction mechanism. It should make it possible to check for thermodynamic consistency of cycles of changes in protonation and tautomer states such as seen if Fig. 1.

SAMPL6 pK_a Challenge type I submissions report the predicted pK_as for transitions between selected pairs of prespecified microstates. This information is far richer than a list of macroscopic pK_as. However, as the list of individual microscopic pK_as were analyzed it became apparent that this format is also far from ideal. A list of pK_as provides only a local view of the relative proton affinity of pairs of states of the molecule. In addition, the list can have more information than is needed. For example, for SM07 (Fig. 1) there are 24 possible pK_as between all pairs of microstates that vary by one proton (adjacent rows), of which 17 are shown. However, we will show that knowing the relative free energy of the 11 individual microstates at a single reference pH (ΔG°) can completely describe the free energy (ΔG) of all microstates at any pH. Populations of each microstate can then be obtained given the ΔGs to recover titration curves. In addition, is not readily apparent if the overall free energy landscape built up from the network of individual pairwise pK_as is thermodynamically consistent, while this is straightforward to see when each microstate of the molecule is associated with its relative free energy.

Deriving ΔG°s for a network of protonation and tautomeric states from a list of pK_as

Describing a molecule with 3 pK_as and no tautomeric states by the relative microstate free energy at pH 0 or 7

Currently, the information we have about the protonation and tautomeric states of protonatable molecules is often collected as a set of pK_as and this was the information submitted to SAMPL6. We will therefore show how to use these pK_as to build up a standard state free energy ladder, which is the free energy differences between microstates at one pH. The first example considers three pK_as separating four states, denoted A, B, C, and D (Table 1). A has three dissociable protons and D none. The analysis would be the same if the input pK_as come from experiment or simulation. Tautomeric microstates, with the same number of protons but at different locations on the molecule, will be considered in the next section. The pK_as are at − 2.17, 5.61, and 13.77. These are taken from the EPIK predictions for the pK_a between the SM07 microstates 14, 7, 4, and 12 (Figs. 1,2) [24]

To determine the relative free energy of the four microstates at a single pH one state is chosen to be the reference state. The reference state and pH are arbitrary choices, but simply need to be applied consistently. We will describe the calculation of the relative state free energies with B as the reference (ΔG°_{jB}) and then show that using C as a reference (ΔG°_{jC}) provides the same relative energies between all states, but with a constant offset equal to the energy difference between microstates B and C (ΔG°_{BC}). The reference state is defined here as the second term in the subscript.

Table 1 The pK_as and number of protons provide the input information needed to describe a molecule with 4 protonation states separated by 3 pK_as

Molecule		XH_3^{3+}		XH_2^{2+}		XH^+	X
Microstate		A		B		C	D
Relative charge		i + 3		i + 2		i + 1	i
pH where $G_j = G_k$	pK_{AB}	− 2.17	pK_{BC}	5.61	pK_{CD}	13.77	
ΔG_{jk}	ΔG°_{AB}	2.17	ΔG°_{CB}	− 5.61	ΔG°_{DC}	− 13.77	
Δm_B		1		0		− 1	− 2
Δm_C		2		1		0	− 1

$pK_{jk} = pK_{kj}$ is the pH where $G_j = G_k$; ΔG°_{jk} is the free energy difference between states j and k at pH 0. The reference state is the second term in the subscript. $\Delta G^\circ_{jk} = -\Delta G^\circ_{kj}$. Δm_j is the number of protons relative to the reference state, which is B or C here. As there are no tautomeric states in this simplified example, the microstates and macrostates are equivalent. Free energy (ΔG) is given as unitless free energies where a one unit change in ΔG yields a tenfold population change

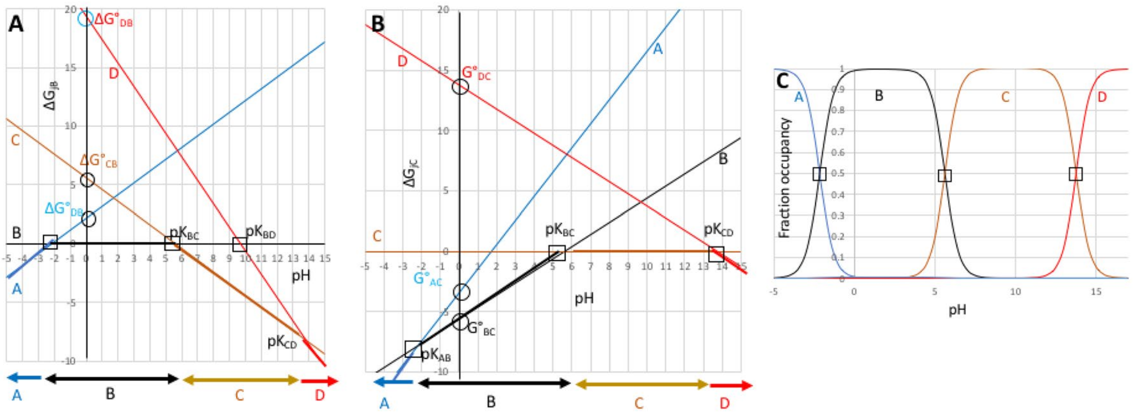

Fig. 2 Relative standard state free energies (ΔG°) and relative number of protons (Δm) is all that is needed to completely describe the pH dependence of a system with multiple protonation states. **a** The relative free energies of the states A,B,C, and D as a function of pH given the pK_as in Table 1 transformed to standard state free energies, ΔG° (Table 2). Relative free energies of microstates change linearly with respect to pH (Eq. 2b). Squares are input pK_as which can be experimentally observable. Circles mark ΔG°_{jB}, the free energies at pH 0 of the other three states relative to B. Horizontal arrows at the bottom show the state at lowest free energy (dominant population) in each pH range. **b** states A, B, C and D with state C as the reference; **c** Titration showing relative state populations vs pH predicted using ΔG°s in Table 2 and Eqs. 3 and 4. This plot is the same independent of which state is used as the reference. ΔG is given in unitless free energies where a unit change in ΔG yields a tenfold population change

If B is the reference then its energy is zero and independent of pH, as shown by the horizontal black line at $\Delta G = 0$ in Fig. 2a. The pK_{AB} at − 2.17 gives the pH where state A and B have equal energy so the line describing the pH dependence of the relative free energy of A crosses B here. The free energy difference between A and B at any other pH is:

$$\Delta G^{pH}_{AB} = \Delta m_{AB} C_{units}(pH - pK_{AB}) \tag{2a}$$

ΔG_{AB} is 0 when the pH is equal to the pK_{AB}. C_{units} moves the values into the desired units of energy. It is 1.36 for kcal/mol or 5.69 for kJ/mol. We will use $C_{units} = 1$, which is $RT\log_{10}10$. Thus, one unit of energy changes the equilibrium constant by a factor of 10 at the reference temperature. A change in pH of 1 unit leads to a 1 unit change in ΔG_{AB} if a proton is gained or − 1 unit if a proton is lost. C_{units} can be referred to as pH units. As A has one more proton than B ($\Delta m_{AB} = 1$) its energy increases with pH as the proton concentration decreases. The standard state reference energy for A is its energy relative to (the reference state) B at the (reference) pH of 0 is:

$$\Delta G^{\circ}_{AB} = \Delta m_{AB} C_{units}(-pK_{AB}) \tag{2b}$$

ΔG°_{AB} is thus the y-intercept in Fig. 2a. In biochemistry, the standard state (ΔG°') is often defined at pH 7 not pH 0. ΔG°' is provided in Table 2, and can be read off Fig. 2a or b from the y value for each state at pH 7. At pH 7 the relative free energies are:

$$\Delta G^{o\prime}_{AB} = \Delta G^{7}_{AB} = \Delta m_{AB} C_{units}(7 - pK_{AB}) \tag{2c}$$

The pK_a at pH 5.61 connects state C to the reference state B. The pK_{BC} is marked on Fig. 2a as the point where the two states have equal energy. As C has one less proton

Table 2 State energies derived from pK_as in Table 1 using different reference states or reference pHs

State j	Δm_{jB}	pH = 0 ΔG°_{jB}	pH = 7 $\Delta G^{\circ\prime}_{jB}$	Δm_{jC}	pH = 0 ΔG°_{jC}	$\Delta\Delta G^{\circ}_{jC}$ − $\Delta\Delta G^{\circ}_{jB}$
A	1	2.17	9.17	2	− 3.44	− 5.61
B	0	0	0	1	− 5.61	− 5.61
C	− 1	5.61	− 1.39	0	0	− 5.61
D	− 2	19.37	5.37	− 1	13.77	− 5.61

With B as the reference state ΔG°_{jB} and ΔG°_{jC} obtained with (Eq. 2b) given pK_{AB} and pK_{CB}; $\Delta G^{\circ}_{DB} = \Delta G^{\circ}_{DC} + \Delta G^{\circ}_{CB}$ (Table 1). $\Delta G^{\circ\prime}_{jB}$, the relative state energies with the reference pH = 7, is obtained with (Eq. 2c). When C is the reference state the ΔG° relative to states B and D are obtained directly from the reported pK_as using (Eq. 2b). $\Delta G^{\circ}_{AC} = \Delta G^{\circ}_{AB} + \Delta G^{\circ}_{BC}$. There is a constant offset for all ΔG°s moving from a system that uses state B or state C as the reference state. ΔG represents unitless free energies where a unit change in ΔG yields a tenfold population change than B ($\Delta m_{CB} = -1$) the free energy of C decreases relative to B with increasing pH, so its energy has a slope of − 1 (in pH units). Extrapolation of the free energy as a function of pH back to the reference pH of 0 yields the $\Delta G°_{CB}$ of 5.61 (Eq. 2b).

While the pK$_a$s in Table 1 give the pairwise free energy difference between states A or C and B, there is no direct information about the transition between states B and D, which differ by 2 protons. The pK_{CD} at 13.77 connects states C and D. Thus, $\Delta G°_{DC} = \Delta m_{DC}(-pK_{DC}) = 13.77$. $\Delta G°_{DB} = \Delta G°_{DC} + \Delta G°_{CB}$ (i.e. the free energy change from B to C plus that from C to D) (Table 2). The slope of ΔG_{DB} is − 2 as D has 2 fewer protons than B. The slope of ΔG_{DC} with pH is − 1, which is not easy to see from the graph, but will become apparent in the next section when C is used as the reference state. At each pH the predominant species will be the state at lowest energy shown by thicker lines in Fig. 2a. Thus, below pH -2.17 this is state A; between − 2.17 and 5.61 it is state B; and above 5.61 it is state C and above pH 13.77 D is the lowest energy and thus the predominant species.

Translating the pK$_a$s, which each connect a pair of microstates, into relative ΔG° for the ensemble of four microstates provides additional information. Thus, the free energy difference between states not connected by a defined pK$_a$, such as a two proton transition between A and D, can be obtained from the sum of stepwise ΔGs at any pH. The crossing points between any pair of lines on the free energy vs pH plot show the pH where two microstates have equal energy and thus equal probability.

The selection of the reference state is arbitrary. Figure 2b shows the graphical analysis of the same pK$_a$s shown in Table 1 but with state C as the reference, instead of B. Now C lies along the horizontal at $\Delta G = 0$. B has one more proton than C so the pH dependence of ΔG_{BC} has a slope of 1 and $\Delta G = 0$ at pK_{BC}. State D has one less proton than C so ΔG_{DC} changes with pH with a slope of − 1. ΔG_{DC} is 0 at pH 13.77, at pK_{DC}. Now it is the pK_{AB} at − 2.17 that is not directly connected to the reference state. $\Delta G°_{AC}$ is $\Delta G°_{AB} + \Delta G°_{BC}$ (Tables 1, 2). The two graphs in Fig. 2a and b are the same except for a rotation to move from B being on the x axis to place C on this axis. For any microstate (j) $\Delta G°_{jB}$ and $\Delta G°_{jC}$ differ by the difference in energy between the states B and C ($\Delta G°_{CB}$), which is − 5.61 at pH 0. As the relative energy difference between all states are the same at each pH the lowest energy (and hence highest population) state at each pH is the same in Fig. 2a and b.

Given the relative energy at pH 0, the relative energy of each state can be determined at any pH by:

$$\Delta G_{jB}^{pH} = \Delta G_{jB}^{\circ} + \Delta m_{jB} C_{units}\left(pH - pH_{ref}\right) \tag{3}$$

Given the energy as a function of pH (ΔG_{jB}^{pH}) the fraction of each state, N_j, at each pH is obtained from the standard expression:

$$N_j^{pH} = \frac{10^{-\Delta G_{jB}^{pH}}}{\sum_i 10^{-\Delta G_{iB}^{pH}}} \tag{4}$$

Plotting N_j^{pH} vs. pH provides the titration curve. The crossing points of titration curves recover the initial, input pK$_a$s (Fig. 2c).

Microstate analysis of SM07

In the microscopic analysis of SAMPL6 pK$_a$ challenge target SM07, eleven microstates were enumerated (Fig. 1). There were 32 blind submissions of microscopic pK$_a$ predictions from eight laboratories [16, 21, 24]. Four research groups submitted a single set of predictions, two submitted 2 distinct sets of predictions, another submitted 10 [19, 25], and a final group 14 [23]. As few as a single pK$_a$ was submitted for this compound (one prediction set) and as many as 17 pK$_a$s were reported (24 prediction sets). We will show how converting all the pairwise pK$_a$s to state ΔG°s will make it easier to compare the entire free energy landscape predicted by different methods, recognizing thermodynamic inconsistencies and ending with better appreciation of whether different calculation methods are converging to similar answers for states that are not experimentally accessible.

Choosing the reference state

We will first consider the values calculated with the program Epik [26–28]. From independent Epik calculations run with the -pH option at pH values between 2–12 (0.1 pH units apart), 8 microstates were predicted to be populated (Fig. 3b; Table 3) [24]. We will then describe the relative microstate energies obtained from the pK$_a$s in all submissions, which describe as many as 11 microstates (Fig. 1).

Knowing the structure of each state we can count the number of protons (Fig. 1). As shown above the choice of reference state is arbitrary. One choice, which is easier to automate, is to use microstate 12 or 16 with the fewest or most protons (Fig. 1). However, we chose state 4 as the reference as it has four reported pK$_a$s. This allows the $\Delta G°_{j4}$ for microstates 6, 7, 14 and 12 to be determined directly with (Eq. 2b); Table 3). $\Delta G°_{14,4}$ is then the sum of $\Delta G°_{14,7} + \Delta G°_{7,4}$.

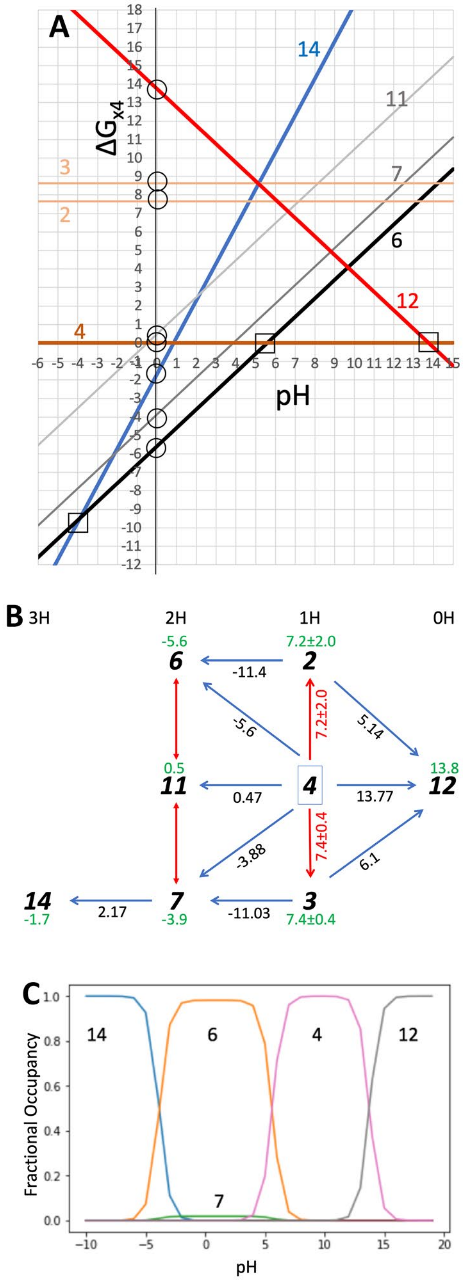

Fig. 3 Graphical depiction of the microstate energy as a function of ▶ pH and resultant titration curve for the eight microstates of SM07 described in Table 3. **a** Graphical representation of the 8 microstate energies as a function of pH using ΔG°s and Δms from Table 3. The squares show pK_as that would be seen experimentally as they connect the states that are at low energy at that pH and the triangles show pK_as that were the input to the calculation. There is an inconsistency in the relative energy of states 2 and 3 calculated when state 12 or state 6 or 7 are used to obtain the free energy difference from reference state 4 (Table 3). **b** The microstate network of 8 microstates of SM07 connected by pK_as calculated with Epik [24]. Microstates predicted by Epik are a subset of those shown in Fig. 1. Dark blue arrows are the ΔG° between the two microstates; Red arrows are ΔG° between tautomers. The standard deviations for $\Delta G°_{2,4}$ and $\Delta G°_{3,4}$ represent the standard error for the free energy calculated around the two nearest closed triangular loops. Green numbers under microstate identifiers are ΔG°, the free energy relative to state 4 at pH = 0.. **c** The probability of each state as a function of pH. Note that while state 6 is the predominant microstate between pH − 5 and 5, a small amount of tautomer microstate 7 is seen. ΔG represents unitless free energies where a unit change in ΔG yields a tenfold population change. Python scripts and interactive Jupyter notebooks to generate networks and graphs of the relative free energy as a function of pH from a list of microscopic pKas can be found at https://github.com/choderalab/titrato

Determining the ΔG° between tautomers uncovers a lack of thermodynamic consistency

Microstates 2, 3, and 4 are tautomers with the same number of protons. Thus, their relative energy is independent of pH, so there is no pH where their energy is equal and thus no pK_a can be defined. However, examination of Fig. 3b shows the ΔG° can be defined between these states by the summed energy along any path to the reference. For example, to determine $\Delta G°_{2,4}$ we consider two short paths: one with state 6 as the intermediate, and one via state 12. The two paths give a different $\Delta G°_{2,4}$ (Table 3). This indicates that the closed reaction cycle from microstate 4 to 6 to 2 to 12 and back to 4 does not sum to zero as it should for a thermodynamically consistent method. The summed free energy for the protonation and tautomer changes are seen to be described as a closed loop when the molecule is described as a network of protonation and tautomeric microstates with energies defined against a single reference state. When different cycles do not sum to zero there are multiple choices for the derived ΔG° values, from using one cycle, averaging the 2 shortest cycles as carried out here, to averaging the results from all possible cycles. When the thermodynamic cycles close properly there is no ambiguity in the relative free energy of the microstates.

Graphical analysis of the microstate energy as a function of pH.

Figure 3a provides a graphical picture of the microstate energy as a function of pH obtained from the pK_as in Table 3. Plotted relative microstate free energy vs. pH shows the energy of each of the two groups of tautomers (1H macrostate, microstates 2, 3, 4) and (2H macrostate, microstates 6, 7, 11) are parallel to each other as the ΔG between them is independent of pH. The graph shows which state(s) are at experimentally accessible energy in any pH range. This is microstate 14 at low pH, a mixture

Table 3 (a) Epik microscopic pK_as for SM07. (b) Epik SM07 microstate standard state free energies

(a)

State$_j$	State$_k$	p$K_{j,k}$	Std err		ΔG°_{jk}
2	12	5.14	1.47	$\Delta G^{\circ}_{12,2}$	5.14
3	12	6.1	0.86	$\Delta G^{\circ}_{12,3}$	6.1
4	12	13.77	0.73	$\Delta G^{\circ}_{12,4}$	13.77
6	2	11.4	2.22	$\Delta G^{\circ}_{6,2}$	-11.44
6	4	5.6	1.13	$\Delta G^{\circ}_{6,4}$	-5.6
7	3	11.03	2.22	$\Delta G^{\circ}_{7,3}$	-11.03
7	4	3.88	2.22	$\Delta G^{\circ}_{7,4}$	-3.88
11	4	-0.47	0.92	$\Delta G^{\circ}_{11,4}$	0.47
14	7	-2.17	2	$\Delta G^{\circ}_{14,7}$	2.17

(b)

State	$\Delta m_{j,4}$	$\Delta G^{\circ}_{j,4}$	Source	$\Delta G^{\circ\prime}_{j,4}$
12	-1	13.77	p$K_{12,4}$	6.77
2	0	5.8 8.63	$\Delta G^{\circ}_{6,4} - \Delta G^{\circ}_{6,2}$ $\Delta G^{\circ}_{12,4} - \Delta G^{\circ}_{12,2}$	5.8 8.63
3	0	7.15 7.67	$\Delta G^{\circ}_{7,4} - \Delta G^{\circ}_{7,3}$ $\Delta G^{\circ}_{12,4} - \Delta G^{\circ}_{12,3}$	7.15 7.67
4	0	0	Reference	0
6	1	-5.6	p$K_{6,4}$	1.4
7	1	-3.88	p$K_{7,4}$	3.12
11	1	0.47	p$K_{11,4}$	7.47
14	2	-1.71	$\Delta G^{\circ}_{14,7} + \Delta G^{\circ}_{7,4}$	12.29

The standard error (Std err) is the value reported by the Epik calculations. Table 3 pK_as are obtained from [24]. Microstate j has one more proton than microstate k. ΔG°_{jk} is the free energy difference between states j and k at pH 0 (Eq. 2b). State IDs are given in Figs. 1 and 3b. Table 3 $\Delta m_{j,4}$ is the change in the number of protons relative to state 4. ΔG°_{j4} and $\Delta G^{\circ\prime}_{j4}$ are the relative microstate free energy at pH 0 and 7 respectively. Source: Where p$K_{j,k}$ is indicated (Eq. 2b and Table 3) are used; otherwise the ΔG°s listed are summed to generate the energy difference between the desired microstate and microstate 4 using the network in Fig. 3b.The energy of microstates 2 and 3 are obtained from the sum of the free energies around the indicated loop using the path described and the resultant ΔG° is the average of the two paths. ΔG presented as unitless free energies where a unit change in ΔG yields a tenfold population change

of 6 and 7, then microstate 4 and at high pH microstate 12 predominates.

The nine pK_as that are given each represent crossing points between two lines on the graph. There are many other possible pK_as that can be read off the graph or obtained by determining the pH where two microstates with different numbers of protons have the same energy. For example, p$K_{7,14}$ is given but p$K_{6,14}$ might be more important as microstate 6 is at lower energy than the tautomer microstate 7. In addition, inconsistencies in the energies obtained for different pairs of pK_as are seen. Thus, the intersection of states 2 and 6 as well as that of 3 and 7 are different than the reported pK_as because the y intercept (ΔG°) represents the average of two thermodynamic cycles while the slope of the lines for 2 and 3 must be fixed at 0 as these microstates have the same number of protons as reference state 4, or the slope is 1 if the microstate has an additional proton (e.g. for 6,7, and 11). The derived titration curve highlights the low energy, experimentally accessible states. It should be noted that the titration covers a pH range that is higher and lower than most experiments. The 2H protonation state is found to be the most stable between pH -5 and 5. While tautomer 6 is dominant, the free energy analysis shows tautomer 7 is close in energy so would be predicted to be a minority species. The relative probability of microstates 6 and 7 is pH independent and their sum gives the probability of the 2H macrostate as a function of pH.

ECRISM-13 (SAMPL6 pK_a challenge submission ID 0xi4b) reported 17 pK_a values for SM07 (Fig. 4) [23]. The network obtained from these pK_as show all closed paths around the network have a summed ΔG° of zero, indicating the reported values are thermodynamically consistent (Fig. 4a). Now there are predictions for the free energy of tautomers 13, 14 and 15 as well as the highly protonated microstate 16. The pattern of relative microstate energies are in qualitative agreement with the Epik simulations and the resulting titration is also similar. The same microstates are at low energy (Fig. 4c). Both calculations place

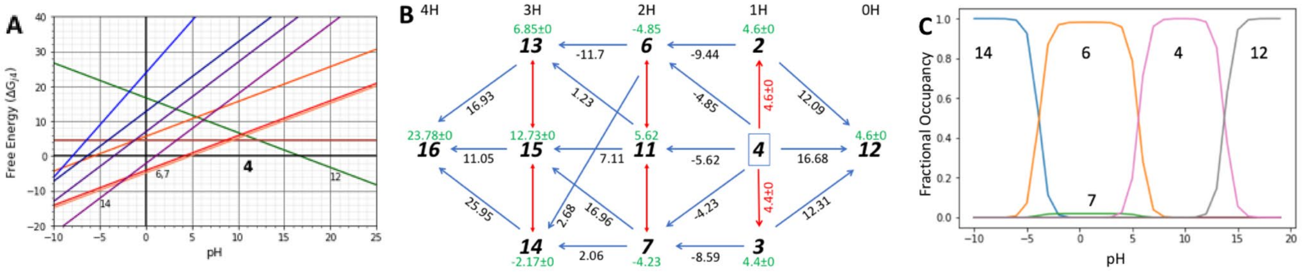

Fig. 4 Graphical analysis of a network of 11 microstates SM07. **a** ΔG_{j4} as a function of pH for all microstates. $\Delta G°_{j4}$ is ΔG_{j4} at pH 0. ΔG represents unitless free energies where a unit change in ΔG yields a tenfold population change. **b** Network of 11 SM07 microstates considered by calculation ECRISM-13 (ID 0xi4b) [23], the resultant pairwise $\Delta G°_{j4}$ (blue arrows) and microstate $\Delta G°_{j4}$ relative to microstate 4 (green numbers). **c** Predicted titration curves given the microstate ΔGs as a function of pH. It should be noted that the pK_as shown in Fig. 4c match the crossing points in Fig. 4b that occur as one lowest energy state is replaced by another as the pH changes

the 2H microstates 6 and 7 close enough in energy that a mixture of tautomers are predicted to be seen (Figs. 3c, 4c).

Overview of all submitted predictions for SM07: Do different calculation methods give similar values for ΔG°?

Table 4 gives the ΔG° for all predictions of all microstates of SM07 with microstate 4 as the reference state and pH 0 the reference pH. It is far more compact than a table of pK_as, requiring only a single value for each microstate. In contrast, a single microstate in the SM07 network of states described in Fig. 2, can be connected by as many as six pK_as to the six microstates that different by one proton. In addition, the pH independent ΔG° between tautomeric states are established, although there is no pK_a that can be defined for a pair of microstates with the same number of protons. As shown in Figs. 2,3 and 4, knowledge of $\Delta G°_{j4}$ at a single pH and the difference in the number of protons from the reference state (Δm_{j4}) can be used to find the microstate free energy differences at all pHs. The analysis shows where any two microstates are at the same energy either graphically or by calculation, and so identifies all possible microscopic pK_as. The pK_as can connect microstates at high energy or that differ by more than one proton. Knowing the microstate energies as a function of pH allows the calculation of their relative probability with pH. Plotting this probability as a function of pH generates a titration curve, visually identifying the low energy states and providing the macroscopic pK_as (Figs. 2,3,4). Table 4, gives the ΔG°s at the reference pH of 0, although the table can be modified for any pH using (Eq. 3) (e.g. Table 2, 3).

Comparison of the calculated results with experiment.

A single experimental pK_a value at pH 6.08 is available for SM07 [5, 17]. SM07 was one of the few whose titration was followed by NMR showing the transition is between microstates 4 and 6. The NHLBI QM submissions (ko8yx, w4z0e, wcvnu, arcko, wexjs) [19] and the Fraczkiewicz submission (hdiyq) as well as the single KirilLanevskij submission (v8qph) predicts both the correct low energy microstates and the pK_a correctly (with a maximum error of 1 pH unit).

The ΔG° analysis gives access to the pH independent ΔΔG between tautomers, which can be compared with the experimental evidence for the transition between the 1H and 2H microstates described by experiment. All calculations put microstate 4 as the lowest energy 1H state, in agreement with experiment. However, the free energy of microstate 7 is often very close to that of 6, so it is often predicted to be a minority species in the titration (Figs. 3c, 4c). It should be noted that in several cases the microscopic pK_as between microstates 7 and 4 is close to the experimental value. However, macroscopic titration will always predominantly involve microstate 6 as it is at lower energy.

The thermodynamic consistency of the submitted predictions for SM07

Viewing the molecule as a network of connected protonation and tautomeric states allows the self-consistency of the relative energies to be determined. If the sum of the ΔG° around a closed path deviates from zero by more than the likely error of the individual values, then this group of microstate energies are not thermodynamically consistent and something is wrong. We can see that different

Table 4 Microstate ΔG°_{i4} for SM07 derived from pK_as submitted to the SAMPL6 Blind pK_a Challenge shows areas of qualitative agreement with significant differences in calculated energies

Microstate														
		0H	1H			2H			3H			4H		
ID		12	2	3	4	6	7	11	13	14	15	16	$\Delta G_{16\text{-}12}$	ΔG_{cycle}
hgn83 [25]	NHLBI-9	3.7	16.7	18.0	0	20.4	28.5	28.5	36.3	24.4	36.5	55.9	52.2	7.9
758j8 [25]	NHLBI-8	8.3	12.4	17.3	0	15.8	23.9	25.6	32.6	28.8	41.1	57.4	49.1	7.9
z3btx [25]	NHLBI-7	9.7	17.0	17.3	0	14.4	23.2	21.8	30.0	18.4	30.2	49.7	40.0	8.3
0wfzo [25]	NHLBI-6	14.3	15.7	16.6	0	9.8	17.2	18.6	26.1	21.8	34.2	50.8	36.5	8.3
8toyp [23]	ECRISM-8	20.6	6.4	5.4	0	– 1.4	– 1.3	4.5	11.6	6.6	19.9	40.7	20.1	0
Arcko [19]	NHLBI-4	16.8	6.9	6.2	0	– $\underline{6.5}$	– 2.7	6.8	12.9	4.7	20.5	36.2	19.4	0.9
Wexjs [19]	NHLBI-5	17.8	6.8	5.9	0	– $\underline{6.0}$	– 2.4	7.5	13.7	5.1	20.5	36.9	19.1	0.3
w4z0e [19]	NHLBI-2	19.0	6.7	5.8	0	– $\underline{5.4}$	– 2.0	8.0	14.2	5.6	20.9	37.0	18.0	0.3
Wcvnu [19]	NHLBI-3	18.1	6.8	6.5	0	– $\underline{5.2}$	– 1.9	7.8	13.3	4.7	19.5	36.0	17.9	0.5
ko8yx [19]	NHLBI-1	18.0	6.9	6.5	0	– $\underline{5.3}$	– 2.2	7.5	13.0	4.5	19.0	35.6	17.6	0.9
z7fhp [23]	ECRISM-7	20.2	6.3	5.4	0	– 1.8	– 1.6	4.0	10.2	5.2	18.4	37.6	17.4	0
Kxztt [23]	ECRISM-1	15.9	5.4	4.1	0	– 2.6	– 2.6	2.0	7.0	3.7	13.3	30.1	14.2	0
Nxaaw [23]	ECRISM-12	17.8	5.5	4.5	0	– 4.8	– 4.4	5.7	8.8	0.4	14.9	30.0	12.2	0
2umai [23]	ECRISM-5	16.7	4.7	4.0	0	– 2.5	– 2.4	1.9	5.7	2.0	11.8	25.7	9.0	0
cm2yq [23]	ECRISM-6	16.6	4.7	4.0	0	– 2.5	– 2.4	1.8	5.7	1.9	11.8	25.6	9.0	0
Epvmk [23]	ECRISM-9	14.0	4.1	3.5	0	– 2.9	– 2.7	2.7	5.5	0.7	11.0	22.8	8.8	0
4o0ia [23]	ECRISM-11	13.8	3.9	3.4	0	– 2.9	– 2.7	2.5	5.3	0.6	10.5	22.1	8.3	0
xnoe0 [23]	ECRISM-10	14.2	4.1	3.5	0	– 3.1	– 2.9	2.5	5.1	0.2	10.5	22.1	7.9	0
ktpj5 [23]	ECRISM-3	16.5	4.7	4.0	0	– 2.7	– 2.6	1.6	4.9	1.1	11.0	24.0	7.5	0
Wuuvc [23]	ECRISM-4	16.4	4.7	4.0	0	– 2.7	– 2.6	1.6	4.8	1.1	10.9	23.8	7.4	0
0xi4b [23]	ECRISM-13	16.7	4.6	4.4	0	– 4.9	– 4.2	5.6	6.9	– 2.2	12.7	23.8	7.1	0
Cywyk [23]	ECRISM-14	16.3	4.5	4.3	0	– 5.0	– 4.4	5.3	6.0	– 2.8	11.7	21.9	5.6	0
ftc8w [23]	ECRISM-2	15.5	4.8	3.7	0	– 4.7	– 4.8	1.2	2.3	– 2.6	7.9	18.5	3.0	0
Gdqeg [23]	PCM-15	8.9	2.9	2.4	0	– 7.2	– $\underline{7.0}$	– 1.4	– 4.5	– 9.4	– 1.1	1.3	-7.6	0
6tvf8 [16]	BannonOE-1	–	–	–	0	– 7.1	– $\underline{5.1}$	– 0.5	– 6.3	– 9.5	– 5.9	– 12.9	NI	8.8
hdiyq	Rfraczkiewicz	–	–	–	0	– $\underline{5.6}$	– 3.7	– 1.3	– 5.4	– 2.8	– 2.1	– 1.2	NI	0
Ccpmw [21]	Wilcken2	15.8	–	–	0	– 7.4	– $\underline{6.0}$	3.4	–	– 7.9	–	–	NI	NI
t8ewk	cosmologic_fine17	–	–	–	0	– 7.4	– $\underline{5.9}$	–	–	–	–	–	NI	NI
v8qph [24]	KirilLanevskij-1	–	–	–	0	– $\underline{5.5}$	–	–	–	–	–	–	NI	NI
	EPIK	13.8	7.2	7.4	0	– $\underline{5.6}$	– 3.9	0.5	–	– 1.7	–	–	NI	

Energy of microstates of SM07 at pH 0, with state 4 used as the reference. ΔG in unitless free energies where a unit change yields a tenfold population change. Microstate IDs as in Fig. 1. The rows are ordered by $\Delta\Delta G^{\circ}_{16\text{–}12}$, which is the energy difference between the unique 4H and 0H microstates. The tautomers, with the same number of protons, are grouped together (1H, 2H, 3H). ΔG°_{j4} values are derived from predicted pK_as found for the SAMPL6 pK_a Challenge: (https://github.com/samplchallenges/SAMPL6/tree/master/physical_properties/pKa/analysis/analysis_of_typeI_predictions/typeI_predictions). The number in the square bracket is the reference and submission ID identifies published papers associated with each submission. The Epik calculations do not have a SAMPL6ID. The only experimentally observed pK_a is 6.08 ($\Delta G^{\circ}_{64} = -6.08$) is obtained by spectrophotometric method which technically measures the macroscopic pK_a for multiprotic molecules. NMR analysis identified microstate state 4 and 6 as dominant microstates that participate in this transition [5]. Values within 1 pH unit of the predicted pK_a are underlined. ΔG_{cycle} is the largest difference in energy calculated for the summed ΔG°s along all cycles of length 4 in the graph of SM07 microscopic equilibria for each of the submissions. NI indicates that too few pK_as were submitted to generate closed cycles to check for thermodynamic consistency

submissions have different degrees of internal constancy. The ΔG° were summed along all cycles of length 4 in the graph of SM07 microscopic equilibria for each of the submissions. Table 4 gives the largest value of the summed ΔG° for each prediction set (ΔG_{cycle}).

The submissions from ECRISM (kxztt, ftc8w, ktpj5, wuuvc, 2umai, cm2yq, z7fhp, 8toyp, epvmk, xnoe0, 4o0ia, nxaaw, 0xi4b, cywyk) [23] have closed free energy cycles, with the exception of rounding errors on the second decimal, as does the Fraczkiewicz submission (hdiyq). The NHLBI QM submissions (ko8yx, w4z0e, wcvnu, arcko, wexjs) [19] have closed cycles for some but not every 4-microstate cycle, but the mismatches are all below 1 ΔpH unit. In contrast, NHLBI submissions using QM-MM (0wfzo, z3btx, 758j8, hgn83) [25], do not produce thermodynamically consistent cycles, with inconsistancies of around 8 ΔpH units for the cycle containing microstates 4, 7, 15, and 11. The submission that used the Bannan OE method (6tvf8) [16] have cycles that do not sum to 0, with the largest cycle error being 8.75 ΔpH units for the cycle between microstates 4, 11, 13, and 6.

It should be noted that the ΔG°_{j4} connects the 1H microstate 4 to other microstates. As described in Table 2b the energy of tautomers is derived from the sum of free energies along a thermodynamic path and the energy of states that are separated from microstate 4 by more than 1 proton (3H and 4H microstates for SM07) are obtained by sums of ΔG°s along the path to the reference state 4. When the cycles do not close than the values in Table 4 become dependent on which path is used or if multiple paths are averaged.

Overview of the SM07 landscapes show qualitative consistency, but large differences in values.

Only one experimental value is available for SM07. Under these circumstances simulations can offer information if the calculations can be vetted in some manner. One check is the ability to match the single known pK_a, identifying the correct macrostates (1H to 2H here) and correct microstates (4 and 6). Another check is that the overall network is thermodynamically consistent. Lastly, we might say that the calculation of pK_as for molecules such as SM07 is 'solved' if the various submissions find similar answers for the relative state energies that can be checked against experiment for the few microstates that are experimentally accessible.

Table 4 allows evaluation of the consistency of the lowest energy microstate at the reference pH. Here this is pH 0, but the table to be remade at any pH (Eq. 3). It is apparent that at pH 0 the calculations do not agree on what is the lowest energy protonation state. It can be the 1H state (NHLBI-6 to 9), the 2H state or a 3 H state. Thus, the calculations do not agree on what is the net charge of the SM07 molecule at the reference pH.

Another comparison amongst the calculations is to compare the ΔG°_{j4} of individual microstates. The overall range of energy from the microstate with no protons (microstate 12) to that with 4 protons (16) varies enormously between the different calculations from − 7.6 for PCM-1 to + 52 ΔpH units for NCBLI-9 at pH 0. The calculations which are not thermodynamically consistent (NHBLI 6,7,8,9 (on the left in Fig. 5) and PCM (on the right) are clearly different from the bulk of the calculations. The thermodynamically consistent networks still have a range of energy from the most to least protonated microstates of 20 to -7.6 ΔpH units. As one unit

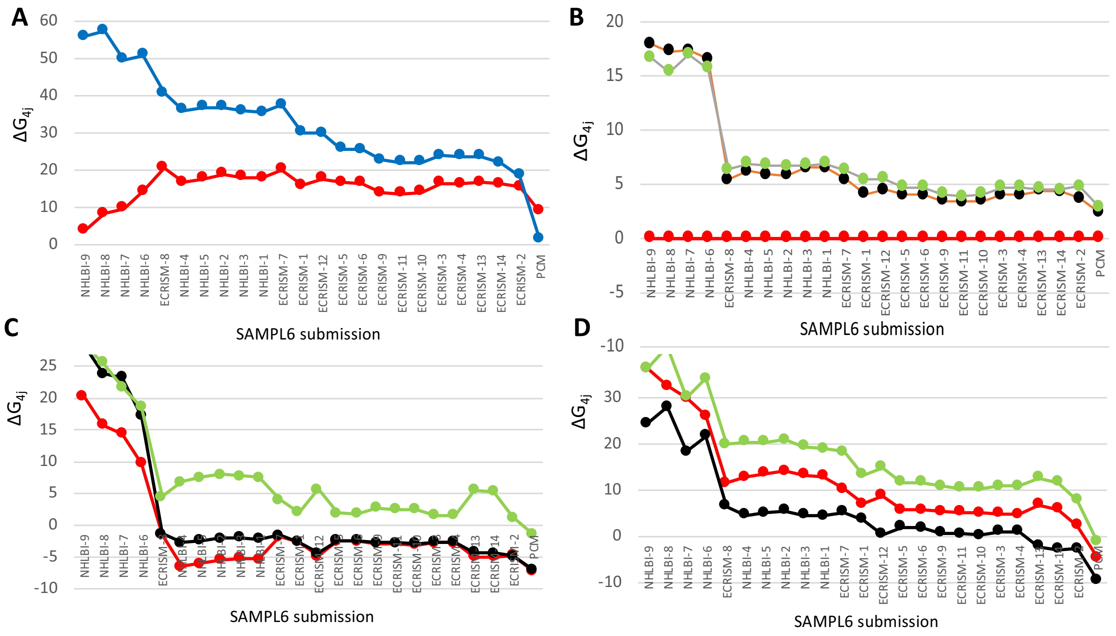

Fig. 5 Overview of relative free energies for individual microstates for individual submissions to the SAMPL6 blind pK_a challenge. All submissions that provided information about all 11 microstates are included, ordered by the free energy difference between the 4H microstate (16) and the 0H microstate (12). Data from Table 4; definition of microstates from Fig. 1. **a** Blue: microstate 16, Red: 12; **b** Green: microstate 2; Black: 3; Red: 4 (reference state); **c** Green: microstate 11; Black: 7; Red: 6; **d** Green: microstate 15; Black: 14; Red: 13

of energy is sufficient to change the relative population by tenfold, this represents a large variation. Thus, while the $\Delta G°_{12,16}$ may not be significant experimentally, the difference in this value shows that the free energy landscape for this molecule is predicted to be radically different in the different calculations. The array of different values shows how SAMPL challenges allow the strengths and weaknesses of different computational to be seen. Outliers have much to teach us.

The relative tautomer energies are, perhaps, a simpler test of the various calculation methods, as the compounds have the same net charge so there is likely to be a smaller difference in solvation energy or influence by the uncertainties in the calculation of the energy of the free proton. The relative energy of each set of tautomers is independent of pH. All calculations show microstates 2 and 3 of SM07 to be close in energy and at higher energy than microstate 4 (Fig. 5b). Likewise, microstates 6 and 7 are generally close in energy with 6 being the lower and state 11 being significantly higher in energy (Fig. 5c). For the tautomers with three protons (microstates 13,14 and 15) microstate 14 is predicted to be the lowest energy microstate, but there is less agreement about the relative energy of these tautomers (Fig. 4d). Thus, overall comparison of the thermodynamically consistent networks of ΔG°s the relative tautomer free energies are in qualitative agreement.

Conclusion

The protonation and tautomer states of extended organic molecules will significantly influence their solubility, partition coefficients and binding affinities to biologically important macromolecules. Molecules used as drugs often have multiple protonation and tautomeric states. Thus, we need to be able to organize the information about the molecular macrostates (with different charge) and microstates (defining the position of all protons) so that under any set of conditions we can determine the dominant charge of the compound and its likely tautomeric state. The question addressed here is how to best organize the information we have about these complex molecules.

The SAMPLE6 Blind pK_a Challenge was the first SAMPL challenge directly focusing on the ability of simulation to predict pK_as of complex organic molecules [5]. Evaluating the submissions made it clear that the best way to describe the pH dependence of molecules with multiple protonation and tautomeric states was not a solved problem. It proved to be difficult to compare different calculations with each other using the complex lists of microstate pK_as. The work presented here shows that reporting only the free energy at a single pH, ΔG°, and change in the number of protons, Δm, each with respect to one (arbitrary) microstate is a better way to report information about protonation and tautomeric states. This procedure should be used for future SAMPL challenges, but it should also be useful as a general way to archive information about molecules with multiple protonation and tautomeric states more generally. It should be noted that this paper shows in detail how to back calculate all the microstate ΔG°s from a list of submitted pK_as. However, computer simulations will often calculate relative microstate ΔG°s, which were then submitted as pK_as.

There are a number of significant advantages to listing microstate ΔG°s for a network of states rather than a list of pairwise pK_a between specific states. The list of ΔG°s is more compact. Thus, for the SM07 microstates considered here, 11 ΔG°s and Δms provide all needed information to determine the free energy difference between any pair of states. In contrast, there are 24 pK_a that connect only states that differ by one proton. The information provided by the ΔG°s is richer. It provides the ΔG°s between tautomers, which is never evident from lists of pK_as as the free energy difference between molecules with the same number of protons is pH independent (Fig. 5b,c,d). The relative energy of microstates that differ by more than 1 proton is clearly defined (Fig. 5a). The summed free energy around closed cycles in the network of microstates can be checked for thermodynamic consistency. As in any equilibrium system, knowledge of the free energy of all states determines the population of each state in the ensemble (Eq. 4). Knowing standard state ΔG°s and that the free energy varies linearly with pH with a slope of that reflects the change in the number of protons relative to the reference state (Δm) directly provides the relative free energy of all states at all pHs (Eq. 3).

The ensemble of all microstate ΔG°s allows the calculations derived by all methods including empirically based methods such as machine learning and QSAR to be compared with each other to determine where all methods agree (Fig. 5). In situations where experimental data is unavailable (and likely to remain so) convergence of values calculated in different ways lends support to the answers obtained by the simulations. The agreement can be qualitative, as often is here, where the ordering of the lowest to highest energy tautomeric state is the same for all calculations. But the numerical free energy differences between states can vary significantly showing that there is more work to be done for these simulation methods to be able to reliably substitute for experimental measurements.

If this round of calculations does lead to a second prediction challenge for pK_as, we would strongly suggest that only microscopic data be reported; that this should be given as the standard state ΔG°; and that only thermodynamically consistent networks of ΔG°s be submitted.

Acknowledgements MRG and TM acknowledge the support of the National Science Foundation grant MCB-1519640. JDC acknowledges support of the National Cancer Institute of the National Institutes of Health under P30CA008748 and partial support from NIH grant P30 CA008748. MI, JDC, and ASR gratefully acknowledge support from NIH grant R01GM124270 supporting the SAMPL Blind Challenges. MI acknowledges support from a Doris J. Hutchinson Fellowship. MI and JDC acknowledge support from the Sloan Kettering Institute and are grateful to OpenEye Scientific for providing a free academic software license for use in this work.

Funding A complete funding history for the Chodera lab can be found at https://choderalab.org/funding

Compliance with ethical standards

Conflict of interest JDC is a member of the Scientific Advisory Board of OpenEye Scientific Software. The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health, the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Bayer, Entasis Therapeutics, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, XtalPi, the Molecular Sciences Software Institute, the Starr Cancer Consortium, the Open systematic Consortium, Cycle for Survival, a Louis V. Gerstner Young Investigator Award, and the Sloan Kettering Institute.

References

1. Martin YC (2009) Let's not forget tautomers. J Comput Aided Mol Des 23(10):693
2. Czodrowski P (2012) Who cares for the protons? Bioorg Med Chem 20(18):5453
3. Seybold PG, Shields GC (2015) Computational estimation of pKa values. Wiley Interdisciplinary Reviews: Computational Molecular Science 5(3):290
4. Fraczkiewicz R, Lobell M, Goller AH, Krenz U, Schoenneis R, Clark RD, Hillisch A (2015) Best of both worlds: combining pharma data and state of the art modeling technology to improve in Silico pKa prediction. J Chem Inf Model 55(2):389
5. Isik M, Levorse D, Rustenburg AS, Ndukwe IE, Wang H, Wang X, Reibarkh M, Martin GE, Makarov AA, Mobley DL, Rhodes T, Chodera JD (2018) pK_a measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments. J Comput Aided Mol Des 32(10):1117
6. Hong J, Hamers RJ, Pedersen JA, Cui Q (2017) A Hybrid Molecular Dynamics/Multiconformer Continuum Electrostatics (MD/MCCE) Approach for the Determination of Surface Charge of Nanomaterials. JPhys ChemC 121:3584
7. Kim J, Mao J, Gunner MR (2005) Are acidic and basic groups in buried proteins predicted to be ionized? J Mol Biol 348:1283
8. Lee AC, Crippen GM (2009) Predicting pKa. J Chem Inf Model 49(9):2013
9. Alexov E, Mehler EL, Baker N, Baptista AM, Huang Y, Milletti F, Nielsen JE, Farrell D, Carstensen T, Olsson MH, Shen JK, Warwicker J, Williams S, Word JM (2011) Progress in the prediction of pKa values in proteins. Proteins: Struct Funct Bioinform 79(12):3260
10. Nielsen JE, Gunner MR, Garcia-Moreno BE (2011) The pKa Cooperative: a collaborative effort to advance structure-based calculations of pKa values and electrostatic effects in proteins. Proteins 79(12):3249
11. Gunner MR, Baker NA (2016) Continuum Electrostatics Approaches to Calculating pKas and Ems in Proteins. Methods Enzymol 578:1
12. Chen Y, Roux B (2015) Constant-pH Hybrid Nonequilibrium Molecular Dynamics-Monte Carlo Simulation Method. J Chem Theory Comput 11(8):3919
13. Damjanovic A, Miller BT, Okur A, Brooks BR (2018) Reservoir pH replica exchange. J Chem Phys 149(7):072321
14. Swails JM, York DM, Roitberg AE (2014) Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. J Chem Theory Comput 10(3):1341
15. Chen W, Morrow BH, Shi C, Shen JK (2014) Recent development and application of constant pH molecular dynamics. Mol Simul 40(10–11):830
16. Bannan CC, Mobley DL, Skillman AG (2018) SAMPL6 challenge results from pK_a predictions based on a general Gaussian process model. J Comput Aided Mol Des 32(10):1165
17. Isik M, Rustenburg AS, Rizzi A, Bannan CC, Gunner MR, Murakami T, Mobley DL, Chodera JD Accuracy of macroscopic and microscopic pK_a predictions of small molecules evalued by the SAMPL6 blind prediction challenge.
18. Selwa E, Kenney IM, Beckstein O, Iorga BI (2018) SAMPL6: calculation of macroscopic pKa values from ab initio quantum mechanical free energies. J Comput Aided Mol Des 32(10):1203
19. Zeng Q, Jones MR, Brooks BR (2018) Absolute and relative pKa predictions via a DFT approach applied to the SAMPL6 blind challenge. J Comput Aided Mol Des 32(10):1179
20. Pickard FC, Konig G, Tofoleanu F, Lee J, Simmonett AC, Shao Y, Ponder JW, Brooks BR (2016) Blind prediction of distribution in the SAMPL5 challenge with QM based protomer and pK a corrections. J Comput Aided Mol Des 30(11):1087
21. Pracht P, Wilcken R, Udvarhelyi A, Rodde S, Grimme S (2018) High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pKa values in the context of the SAMPL6 challenge. J Comput Aided Mol Des 32(10):1139
22. Tielker N, Eberlein L, Chodun C, Gussregen S, Kast SM (2019) pKa calculations for tautomerizable and conformationally flexible molecules: partition function vs. state transition approach. J Mol Model 25(5):139
23. Tielker N, Eberlein L, Gussregen S, Kast SM (2018) The SAMPL6 challenge on predicting aqueous pKa values from EC-RISM theory. J Comput Aided Mol Des 32(10):1151
24. Rustenburg AS, Isik M, Grinaway PB, Rizzi A, Gunner MR, Chodera JS Predicting small-molecule pK_a values and titration curves for teh SAMPL6 pK_a challenge using Epik and Juguar.
25. Prasad S, Huang J, Zeng Q, Brooks BR (2018) An explicit-solvent hybrid QM and MM approach for predicting pKa of small molecules in SAMPL6 challenge. J Comput Aided Mol Des 32(10):1191
26. Epik SR (2017) 2017–4: Schrödinger. New York, NY, LLC
27. Greenwood JR, Calkins D, Sullivan AP, Shelley JC (2010) Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. J Comput Aided Mol Des 24(6–7):591
28. Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M (2007) Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. J Comput Aided Mol Des 21(12):681