2017

# Model for Screened, Charge-Regulated Electrostatics of an Eye Lens Protein: Bovine GammaB-Crystallin

Christopher W. Wahle
*Rochester Institute of Technology*

K. Michael Martini
*Rochester Institute of Technology*

Dawn M. Hollenbeck
*Rochester Institute of Technology*

Andreas Langner
*Rochester Institute of Technology*

David S. Ross
*Rochester Institute of Technology*

*See next page for additional authors*

## Recommended Citation

Wahle, Christopher W.; Martini, K. Michael; Hollenbeck, Dawn M.; Langner, Andreas; Ross, David S.; Hamilton, John F.; and Thurston, George M., "Model for Screened, Charge-Regulated Electrostatics of an Eye Lens Protein: Bovine GammaB-Crystallin" (2017). *Physical Review E*, 96 (3), 1-25. Accessed from https://repository.rit.edu/article/1872

## Authors

Christopher W. Wahle, K. Michael Martini, Dawn M. Hollenbeck, Andreas Langner, David S. Ross, John F. Hamilton, and George M. Thurston

# Model for screened, charge-regulated electrostatics of an eye lens protein: Bovine gammaB-crystallin

Christopher W. Wahle,[1] K. Michael Martini,[2,3] Dawn M. Hollenbeck,[2] Andreas Langner,[4,*] David S. Ross,[1]
John F. Hamilton,[1] and George M. Thurston[2,†]

[1]*School of Mathematical Sciences, Rochester Institute of Technology, Rochester, New York 14623, USA*
[2]*School of Physics and Astronomy, Rochester Institute of Technology, Rochester, New York 14623, USA*
[3]*Department of Physics, University of Illinois at Urbana-Champaign, Urbana-Champaign, Illinois 61801, USA*
[4]*School of Chemistry and Materials Science, Rochester Institute of Technology, Rochester, New York 14623, USA*

(Received 19 October 2016; revised manuscript received 21 July 2017; published 25 September 2017)

We model screened, site-specific charge regulation of the eye lens protein bovine gammaB-crystallin ($\gamma$B) and study the probability distributions of its proton occupancy patterns. Using a simplified dielectric model, we solve the linearized Poisson-Boltzmann equation to calculate a $54 \times 54$ work-of-charging matrix, each entry being the modeled voltage at a given titratable site, due to an elementary charge at another site. The matrix quantifies interactions within patches of sites, including $\gamma$B charge pairs. We model intrinsic $p$K values that would occur hypothetically in the absence of other charges, with use of experimental data on the dependence of $p$K values on aqueous solution conditions, the dielectric model, and literature values. We use Monte Carlo simulations to calculate a model grand-canonical partition function that incorporates both the work-of-charging and the intrinsic $p$K values for isolated $\gamma$B molecules and we calculate the probabilities of leading proton occupancy configurations, for $4 < p\text{H} < 8$ and Debye screening lengths from 6 to 20 Å. We select the interior dielectric value to model $\gamma$B titration data. At $p$H 7.1 and Debye length 6.0 Å, on a given $\gamma$B molecule the predicted top occupancy pattern is present nearly 20% of the time, and 90% of the time one or another of the first 100 patterns will be present. Many of these occupancy patterns differ in net charge sign as well as in surface voltage profile. We illustrate how charge pattern probabilities deviate from the multinomial distribution that would result from use of effective $p$K values alone and estimate the extents to which $\gamma$B charge pattern distributions broaden at lower $p$H and narrow as ionic strength is lowered. These results suggest that for accurate modeling of orientation-dependent $\gamma$B-$\gamma$B interactions, consideration of numerous pairs of proton occupancy patterns will be needed.

DOI: 10.1103/PhysRevE.96.032415

## I. INTRODUCTION

In considering interactions between proteins in solution, one interesting feature is that many different protonation patterns of the titratable, possibly charged amino acid residues coexist in equilibrium [1–3]. Because its acidic and basic residues continually exchange protons with the surrounding solution, an individual protein molecule presents many different spatial patterns of positive and negative charges to its neighbors, and the corresponding voltage patterns around each molecule keep changing. Each possible pair of such charging patterns can in principle give rise to a distinct spatial and orientational dependence of the screened electrostatic interaction between two nearby protein molecules [4–6], and the basins of attraction and repulsive parts of the corresponding potential energy landscape may change in depth or height, angular and spatial extent, and number.

The probabilities of individual charging patterns on molecules that are close enough, approximately within one or two Debye electrostatic screening lengths, also change in response to the altered voltages at neighboring sites on the two surfaces [2,7–11]. Such proximity can already occur more than 20% of the time even at protein volume fractions near 1% [12] and is of critical importance at the high macromolecular volume fractions in living cells, which have been estimated to range from 0.07 to 0.40 [13].

Phase transitions are ubiquitous in the normal and pathological physiology of living cells and tissues [14–19]. Many of these transitions involve multiple chemical equilibria, such as the protonation equilibria studied here. Such protonation and other ligand-binding features, in solutions of proteins and other macromolecules that undergo phase transitions, are analogous to the simultaneous multiple chemical equilibria and phase transitions occurring in micellar solutions, microemulsions, and other self-associating systems [20–24]. How do the relevant chemical equilibria and kinetics affect phase transitions in macromolecular solutions?

For a protein with 20 residues that may change charge at a particular $p$H, by accepting or donating protons from its surroundings, there are $2^{20}$ or about $10^6$ such coexisting protonation patterns. Even if most of these patterns are highly unlikely, it still may be necessary to consider the interactions of many different pairs of charging patterns in order to build quantitative models of their consequences for protein-protein interactions. For example, within a given pair of $\gamma$B molecules at $p$H 7.1, the present model predicts that on each molecule, one or more of the most frequent 100 charging patterns will be present about 90% of the time. Thus, in order to account for 80% of the possible pair interaction potentials between these molecules, in principle one then needs to consider the approximately 5050 distinct pairings that can occur between members of these top 100 charging patterns.

[*] Deceased.

[†] Author to whom correspondence should be addressed: georgemthurston@gmail.com

Therefore, one important element for understanding protein-protein interactions is to know how often each charge pattern occurs in the isolated molecules, which is the focus of the present work. While the probabilities of each of these charge patterns will change with increasing protein concentration, their probability distributions for the isolated protein molecules nevertheless form part of the groundwork for characterizing pairwise and higher-order orientation-dependent interactions between proteins.

In order to evaluate how often a given pattern occurs, it is important to account for the fact that substantial electrostatic coupling can occur between charge patterns on a single protein, as well as between neighboring proteins, in the phenomenon known as charge regulation. Due to these electrostatic couplings, the probability of a given pattern is not, in principle, given by the product of the probabilities for each titratable residue to be occupied with a proton. Indeed, on a lattice such couplings can give rise to a charge-patterning phase transition [25].

That is, knowledge of the individual $p$H values at which each titratable residue is occupied with a proton half the time on average, called the $p\mathrm{K}_{1/2}$ values, in combination with the Henderson-Hasselbalch dependence [26] of occupancy on $p\mathrm{H} - p\mathrm{K}_{1/2}$, is in principle not sufficient to evaluate the pattern probabilities. A better description is that effective $p$K values for given groups change in response to neighboring charges [27–30]. However, because neighboring titratable site occupancies can be substantially altered [31–34] from the Henderson-Hasselbalch form, a more comprehensive description can be given by a grand-canonical distribution model or equivalent consideration [2,7,10,11,25,29,32,33,35,36] that incorporates screened electrostatic couplings, as we pursue here for $\gamma$B-crystallin (Protein Data Bank ID 1AMM).

In the present model we calculate a work-of-charging matrix that models screened electrostatic links between titratable sites on the protein, which are assumed to be fixed in position relative to the protein. In so doing it is important to recognize that other factors can contribute that we do not incorporate, including changes in conformation that are important in allosteric effects and in calculations that use more microscopic representations of dielectric properties [31,37–42], hydration and the hydrophobic effect [43], hydrogen-bonding [44,45], static dipole potentials [45–47], and ion binding [48], each of which can also be expected to produce changes in local charge patterns. We note that $\gamma$B-crystallin is believed to have a fairly robust internal structure; for example, circular dichroism measurements [49] showed no significant spectroscopic changes between $-20\,^\circ$C and $60\,^\circ$C, though this does not rule out the possible role of conformational flexibility in affecting the present model. In the larger context of protein-protein interactions, we note that the work-of-charging matrix *also* involves sites on neighboring proteins and itself depends on the relative positions and orientations of the protein neighbors [10].

We focus the present model on studying the probability distributions of the protonation patterns of an eye lens protein, bovine $\gamma$B-crystallin ($\gamma$B). In aqueous solution, the eye lens $\gamma$-crystallins show liquid-liquid phase separation with an upper consolute temperature [12,50–53], a phenomenon that can compromise transparency of the eye lens and has been linked to cataract disease [54]. The human counterpart of bovine $\gamma$B-crystallin, human $\gamma$D-crystallin (HGD), exhibits many single amino acid mutations that lead to congenital cataracts. The effects of a number of these mutations on the phase diagrams of HGD or HGD–$\alpha$-crystallin mixtures are consistent with cataractogenesis [55–60]. These findings motivate the present work, as an aid to building models of how particular amino acid changes affect protein interactions, the resulting phase diagram, and ultimately lens transparency and the cataract.

We build our model for the probability distributions using the notation of a previous paper [10], though it is important to note that study of protein charge regulation has a very long history [1,2,7,27,29,32,33,35,36,61–63] (for a recent review see Ref. [3]) and there are other equivalent sets of notation. Briefly, we use a linearized Poisson-Boltzmann equation to compute a work-of-charging matrix. This matrix enables modeling of the work required to assemble a given pattern of charges on the protein. The linearized Poisson-Boltzmann equation is a useful starting point for this purpose because its linearity allows for the use of superposition in considering the effects of many charges and the work of charging is a symmetric quadratic form in the vectors of site charges [10]. We model the $p$K values of the titratable residues, with a simplified consideration of their dielectric environments. The combined work-of-charging matrix and $p$K values enter into a grand-canonical distribution that models the relative probabilities of occupancy patterns. We fix an assumed constant interior dielectric value through comparison with existing experimental charge vs $p$H data for $\gamma$B. We then use Monte Carlo simulations and direct calculations to study the resulting probability distributions of protonation patterns. We note that while numerical Poisson-Boltzmann solvers are available that provide calculations of the screened electrostatic environment around proteins and other biological macromolecules (see, e.g., Refs. [64–67]) and corresponding acid-base titration characteristics [68], we developed a program with a view towards flexibility in analyzing model systems [10], including the protonation pattern probability distributions studied here, and for ongoing work on protein-protein interactions.

Figure 1 sets the stage for this work, by depicting the screened electrostatic potential that corresponds to the top proton occupancy pattern near neutral $p$K, modeled to occur about 17% of the time. Interestingly, in view of the attractive interactions that lead to liquid-liquid phase separation of this protein [50], the contours of zero voltage extend fairly far from the protein, in comparison with the Debye length, here 6 Å. For a 1:1 electrolyte in water at 298 K, a Debye length of 6 Å corresponds to an ionic strength of 257 mM, close to that at which the $\gamma$B phase diagram has been studied [12,50,52,69–72]. One might expect that negative and positive patches on neighboring molecules, separated by one or more Debye lengths, are quite capable of creating attractions by facing one another at relatively specific orientations.

To help study the resulting voltage variation, Fig. 1(b) shows the sign of the potential on auxiliary spheres that were placed over the top and bottom portions of the molecule for this purpose, about a one-half Debye length from the surface of the protein, with projected positions of possibly charged residues also indicated. Figure 1(c) shows conjoined Lambert azimuthal equal-area projections of the top and bottom auxiliary spheres,

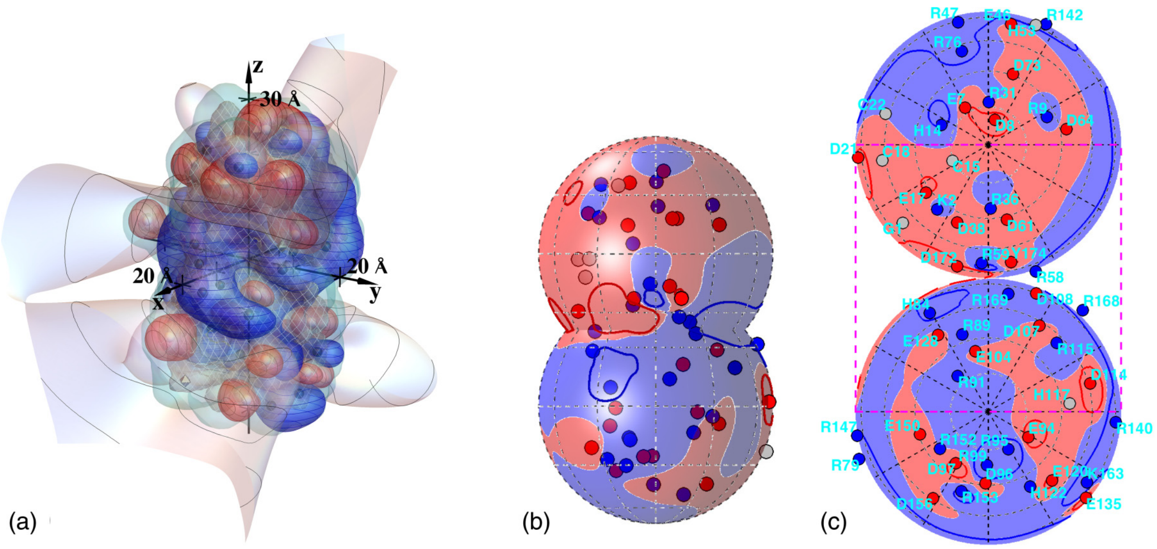

FIG. 1. (a) Screened potential contours produced by the charges of $\gamma$B-crystallin, for the most common protonation pattern occurring at $p$H $= 7.1$ and Debye length 6.0 Å, corresponding to an ionic strength of 257 mM for a 1:1 electrolyte in water at 298 K: $+k_BT/e$ V (blue with horizontal curves), 0 V (light), and $-k_BT/e$ V (red with vertical curves). Black spheres, gray octahedra, and white spheres show positive, neutral, and negative sites, respectively. Curves on the 0 V contour are spaced by two Debye lengths from the center of the protein. The netted surface is the low-dielectric boundary and the plain (light blue) surface just outside it is the electrolyte boundary (see Fig. 2). (b) To aid in visualization, auxiliary spheres of radius 18.5 Å were placed over top and bottom parts of the molecule, and the potential and charges are shown in blue (+) (dark gray and nearly black, respectively) or red (−) (potential light and charges dark gray). Neutral charges are lightest in (b). (c) A simultaneous view of voltages around the entire protein surface can be given with the use of two Lambert azimuthal equal-area projections, one for each of the top and bottom spheres; projected locations of amino-acid residues of possibly charged sites are indicated. The grayscale description of (c) is like that in (b). The dashed rectangle in (c) shows the portion that is visible in (b). The top and bottom perimeter circles in (c) are both images of the crease between the auxiliary spheres in (b). Darker curves in (b) and (c) show $+k_BT/e$ and $-k_BT/e$ V contours.

which provide a single view of the voltages around the entire protein surface.

Thus, the electrostatic interactions between $\gamma$B molecules can contribute to the short-range, orientation-dependent interactions long known to be important for understanding the broad widths of $\gamma$-crystallin liquid-liquid coexistence curves and the position of the crystal solubility boundary, or liquidus [52]. However, because the voltage patterns depend on which $\gamma$B residues are protonated, different protonation patterns may significantly affect the relative orientations that lead to attraction and repulsion, much like the problems that can occur in attempting to fit jigsaw pieces together. Our purpose here is to build groundwork for studying the *distribution* of orientation-dependent interactions that result from probable protonation patterns.

The paper is organized as follows. We briefly recap the relevant theory as it is presented in [10,25]. We then describe the construction of our simplified dielectric model for $\gamma$B, and model $p$K values that would occur in the hypothetical absence of electrostatic interactions between sites, termed the intrinsic, or $p\mathrm{K}_{\mathrm{int}}$, values, as defined by Tanford and Kirkwood [2]. The $p\mathrm{K}_{\mathrm{int}}$ values are functions of the geometry of the interior dielectric environment, because of its effect on the energy stored in the electrostatic field. We then calculate the work-of-charging matrix as a function of the electrostatic screening length in the solvent and the internal dielectric environment, as input to the model grand-canonical partition function. The resulting function gives predictions for the charge vs $p$H, or titration, curve of the protein, which we compare with existing titration [73] and isoelectric point [74,75] data. Because the predictions depend on the assumed internal dielectric coefficient, we make use of the data for tuning this coefficient. We then quantify and study the resulting probability distributions of protonation patterns. Although very few protonation patterns occur compared to the possible ones, we find that their probability distributions are nevertheless broad. We study the extent to which charge pattern probabilities deviate from the multinomial distribution that would result from use of variously defined effective $p$K values, called $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ below, and study how the distributions broaden at lower $p$H and narrow at lower ionic strength. It turned out, somewhat to our surprise, that seemingly subtle changes in the work-of-charging matrix, for example, ignoring entries smaller than $0.2k_BT$, can still produce changes in the modeled rank order of protonation patterns and we analyze why this is so. We briefly discuss possible implications for protein interactions and refinements before concluding.

## II. MODEL

### A. Screened electrostatic model

As in previous work [10,25], we model the response of the electrostatic potential $\phi(\mathbf{r})$ to a specified distribution of fixed charge per unit volume $\rho(\mathbf{r})$ through use of the linearized Poisson-Boltzmann equation [76], written here for a medium with spatially varying relative dielectric coefficient $\varepsilon_r(\mathbf{r})$ and Debye screening parameter $\widetilde{\kappa}(\mathbf{r})$:

$$\nabla \cdot [\varepsilon_0 \varepsilon_r(\mathbf{r}) \nabla \phi(\mathbf{r})] = \widetilde{\kappa}^2(\mathbf{r})\phi(\mathbf{r}) - \rho(\mathbf{r}). \qquad (1)$$

In Eq. (1), $\varepsilon_0$ is the vacuum permittivity, $\varepsilon_r(\mathbf{r})$ is the local, relative static dielectric coefficient, $\phi(\mathbf{r})$ is the local electrostatic potential, $\rho(\mathbf{r})$ is the local free charge per unit volume, and $\widetilde{\kappa}$ is related to the standard Debye screening length in water $1/\kappa$ by $\kappa = \widetilde{\kappa}/\sqrt{\epsilon_w}$, where $\epsilon_w$ is the static dielectric coefficient of liquid water. More sophisticated models of electrolyte solutions are needed in order to accurately model ionic solutions that are not dilute or contain divalent ions and explicit solvent [76–80], to incorporate important physical effects such as finite ion size and ion-specific interactions, including ion absorption [81–83], to include dipolar and polarizability-related interactions [84,85], and to take account of nonlinear dielectric response [86,87]. In the case of ion absorption, for example, one could construct expanded grand-canonical distribution models that would incorporate equilibria with ions other than protons or with polar ligands (see, e.g., [31,88]). Also, the application of Eq. (1) to molecular length scales, on which the protein and the solvent are heterogeneous, involves inherent problems that call for the use of more microscopic, quantum-mechanical approaches, as have been studied for many years (see, e.g., [37,40,89–93] and references therein). Nevertheless, at low ionic strengths and surface charge densities [94], Eq. (1) is a useful starting point for investigating patterned, charge regulation-mediated electrostatic interactions, because its linearity allows the use of superposition in considering the effects of many charges. Also, the work of charging a given configuration of titratable sites may be expressed as a symmetric quadratic form in the vectors of site charges [10].

### B. Grand canonical partition function

In the present model [10], the grand-canonical partition function $\mathcal{Q}$ can be written formally as a sum over the occupancy patterns, indexed by $\alpha$, of protons on the protein

$$\mathcal{Q} = \sum_\alpha e^{-\Delta G_\alpha/k_BT}$$
$$= \sum_\alpha \zeta^{k_\alpha} e^{-(\boldsymbol{\Delta\mu}^0\cdot\mathbf{O}_\alpha)/k_BT} e^{-W_{\mathrm{el},\alpha}/k_BT} \qquad (2)$$

in which $\Delta G_\alpha$ is the free energy of formation of pattern $\alpha$, $\zeta = 10^{-\mathrm{p}H}$, $k_\alpha$ is the total number of protons bound to the protein in configuration $\alpha$, and $\boldsymbol{\Delta\mu}^0 = (\Delta\mu_1^0, \Delta\mu_2^0, \ldots, \Delta\mu_N^0)$ is a vector of standard chemical potential differences for the occupancy of each site. Each $\Delta\mu_i^0$ is related to the corresponding intrinsic $p\mathrm{K}_{\mathrm{int},i}$ value of a titratable site by

$$\exp\left(\frac{\Delta\mu_i^0}{k_BT}\right) = 10^{-p\mathrm{K}_{\mathrm{int},i}}. \qquad (3)$$

By the $p\mathrm{K}_{\mathrm{int},i}$ value we mean the value of the $p$K that site $i$ would have hypothetically in the absence of electrostatic interactions with charges on other sites and in the absence of the electrolytes in the solvent, as in Ref. [2]. That is, it is *not* the $p\mathrm{K}_{1/2}$ that would be measured as the value of the $p$H at which that amino acid residue is, on average, 50% occupied, for example, with use of appropriate nuclear magnetic resonance (NMR) experiments. Instead, $p\mathrm{K}_{1/2}$ values emerge as a consequence of models of the present type [2,10,29,32–34]. In Sec. II D below we describe the model we used for estimating the $p\mathrm{K}_{\mathrm{int},i}$ values.

The vector $\mathbf{O}_\alpha$ in Eq. (2) is the occupancy pattern in configuration $\alpha$, for example, $\{1,0,0,1,1,0,0,\ldots\}$. The quantity $W_{\mathrm{el},\alpha}$ in Eq. (2) denotes the work of charging contribution to the free energy when the protein assumes occupancy pattern $\alpha$. The $W_{\mathrm{el},\alpha}$ is a quadratic form constructed from the work-of-charging matrix $W$, which in this formulation is dimensionless. Each entry $W_{ij}$ in $W$ is the screened electrostatic potential produced at site $i$ by a unit charge at site $j$, multiplied by the electronic charge $e$, and divided by $k_BT$. That $W_{ij} = W_{ji}$ can be shown with use of Eq. (1) [10]. In this notation, $W_{\mathrm{el},\alpha}$ is given by

$$\frac{W_{\mathrm{el},\alpha}}{k_BT} = \frac{1}{2}(q_1,q_2,\ldots,q_n)_\alpha \cdot W \cdot (q_1,q_2,\ldots,q_n)_\alpha$$
$$= \frac{1}{2}(\mathbf{Q}_b + \mathbf{O}_\alpha)\cdot W \cdot (\mathbf{Q}_b + \mathbf{O}_\alpha). \qquad (4)$$

Here the vector $(q_1,q_2,\ldots,q_n)_\alpha$ denotes the actual signed charge numbers on the protein for a specific pattern $\alpha$ and $\mathbf{Q}_b$ denotes the vector of signed, bare charge numbers of the titratable groups, for example, $-1$ or 0. The bare charge numbers are 0 for arginine, histidine, and lysine residues, as well as the terminal amino group, and $-1$ for aspartate, glutamate, cysteine, and the terminal carboxylate. The probability of occupancy pattern $\alpha$, $P_\alpha(\mathbf{x})$, is given by

$$P_\alpha = \frac{e^{-\Delta G_\alpha/k_BT}}{\mathcal{Q}}. \qquad (5)$$

We note that the grand-canonical partition function $\mathcal{Q}$ is also called the binding polynomial [31], because it can be written as a polynomial in powers of the proton activity $\zeta$, as in Eq. (2).

### C. Interior dielectric model and salt exclusion zone

We now describe our model for the quantities $\varepsilon_r(\mathbf{r})$ and $\widetilde{\kappa}^2(\mathbf{r})$ that appear in Eq. (1). We use a simplified model in which $\varepsilon_r(\mathbf{r})$ is assumed to be a scalar that takes a low and constant value inside the protein and a high value outside. After constructing the grand-canonical distribution, we adjusted the interior dielectric coefficient so as to best match the available experimental protein net charge vs $p$H titration data [73,75], as described below in Sec. II F. The value that gave the best match to these data was $\varepsilon_{r,\mathrm{in}} = 3.0$. Outside the protein, we take a value experimentally determined for water at 25 °C, $\varepsilon_{r,\mathrm{out}} = 78.5$ [95]. We modeled the boundary of the low dielectric to be a surface that is 1.4 Å outside the Protein Data Bank (PDB) coordinates of the appropriate atoms (PDB entry 1AMM, from Ref. [96]) and a salt-exclusion zone to extend 3.3 Å beyond this boundary, in approximate accord with hydrated radii of monovalent ions in aqueous solvent [97]. The resulting surfaces are illustrated in Fig. 2.

### D. Model for $p\mathrm{K}_{\mathrm{int}}$ values

In view of our primary present purpose of studying the nature of the probability distributions of the protonation patterns on $\gamma$B-crystallin, we adopted a simple classical approach to modeling the $p\mathrm{K}_{\mathrm{int}}$ values. We start from tabulated $p$K values in water for relevant charged groups [98–100] and then

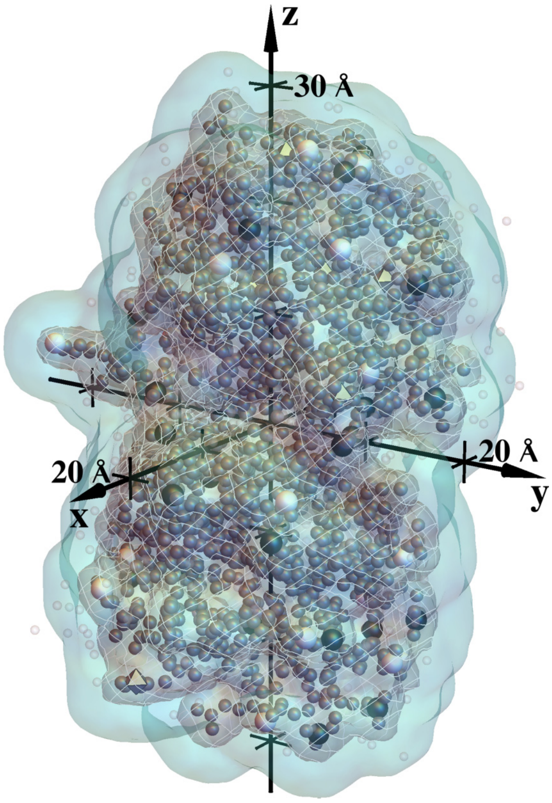

FIG. 2. Illustration of the dielectric and salt-exclusion zone model for bovine $\gamma$B-crystallin, based on PDB entry 1AMM [96], rotated and translated to the coordinate system used for numerical solution of Eq. (1). The dark gray netted surface is the boundary of the low-dielectric region and the light plain surface is the boundary of the salt-exclusion zone, described in the text. Larger black and white spheres and gray octahedra show titratable sites that are positive, negative, and neutral, respectively, for the most probable configuration at $p$H 7.0, modeled to occur about 20% of the time [see Fig. 13(a)]. Smaller dark gray spheres are locations of nonhydrogen atoms in the 1AMM structure and smaller lighter spheres are positions of heteroatoms.

calculate the *change* in the integral that, in a linear dielectric, gives the free energy stored in the electrostatic field per unit volume [101,102], $(1/2)\mathbf{D}(\mathbf{r})\cdot\mathbf{E}(\mathbf{r})$. The integral is taken over the volume outside spheres of radii $r_0$ surrounding the group in question, when water is replaced by a heterogeneous dielectric environment like that near the surface of the protein. This approach omits a number of factors that also affect $p$K values, many of which call for molecular mechanics and/or quantum mechanical treatment [37,39,93,103–107]. These include hydrogen bonding [44,45,108], bound ions, and nonlinear dielectric effects [86,87] that require a different integration of $\mathbf{E}(\mathbf{r})\cdot\delta\mathbf{D}(\mathbf{r})$ than that which yields $(1/2)\mathbf{D}(\mathbf{r})\cdot\mathbf{E}(\mathbf{r})$ [86,101,102]. Hydrogen bonds can stabilize charged carboxylates [44], among other effects, and the presence of metal or other ions, often bound between titratable residues [109], would call for a grand-canonical formulation that involves more exchangeable components [88]. These phenomena are not modeled here. In addition, there are problems involved in characterizing dielectric response at molecular length scales [37,40,89–93], as mentioned above. A related factor not modeled here is the local electrostatic potential from strong static dipoles such as backbone and side-chain amide groups [45–47,110]. There are also solvent effects that can be studied with liquid-state theory approaches [111–113]. However, the present approach is useful as a first approximation; its value may be illustrated, for example, by its remarkable ability to help understand the dependence of salt solubilities on a solvent's static dielectric coefficient [97]. The resulting modeled contribution to the change in $p$K, $\Delta p$K, can be written as

$$\begin{aligned}\Delta pK\,\ln(10) &= \pm\frac{1}{2k_BT}\iiint_{r>r_0}\mathbf{D}(\mathbf{r})\cdot\mathbf{E}(\mathbf{r})dV\\ &\quad-\frac{q^2}{8\pi\varepsilon_0\varepsilon_w r_0 k_BT}\\ (\text{if in uniform solvent}) &= \pm\frac{q^2}{8\pi\varepsilon_0 r_0 k_BT}\left(\frac{1}{\varepsilon_r}-\frac{1}{\varepsilon_w}\right).\end{aligned}\tag{6}$$

In Eq. (6), the + sign is appropriate for groups that become charged when they are *not* occupied by a proton, that is, for glutamate, aspartate, cysteine, and the terminal carboxylate, while the − sign is appropriate for groups that become charged when they *are* occupied by a proton, that is, for lysine, arginine, histidine, and the terminal amino group. To evaluate the needed integral in Eq. (6), as described in detail in the Appendix, we took advantage of the fact that a rough approximation to the shape of our more complicated dielectric model can be constructed by conjoining two spheres, each of radius 15.5 Å. Then we used Kirkwood's analytical solution for the potential due to a charge placed in a low-dielectric sphere, near its surface [61]. We placed the charges a depth of 1.4 Å inside this sphere. We used Gauss's theorem to convert the volume integral of $(1/2)\mathbf{D}\cdot\mathbf{E}$ in Eq. (6) to a surface integral over the small sphere of radius $r_0$; symmetry then permits the needed integral to be converted to a one-dimensional integral, also given in the Appendix, that we evaluated numerically.

To estimate appropriate values of an effective $r_0$ for use in Eq. (6), we used a facility within the quantum-chemistry package GAUSSIAN09 that provides for estimating recommended radii for self-consistent reaction field calculations [114]. For each titratable side-chain group and for the terminal amino and carboxyl groups, we constructed the test molecules listed in Table I, which included the side-chain titratable group in its charged form, and calculated the $r_0$ values in Table I from repeated runs, for which the test ions were surrounded by a medium having the static dielectric coefficient of water. We first used Hartree-Fock calculations with the 6-31G(d,p) basis set to optimize the test molecules in the presence of implicit solvent. Vibration frequency analyses were performed on the optimized structures to determine whether they represented true minima. No structure exhibited imaginary frequencies. We did not perform a conformational analysis of the test molecules or optimize them in their protein environment, for simplicity and consistent with the fact that the present model does not incorporate cross-talk between conformational changes and charge regulation, as noted in the Introduction. We performed $r_0$ calculations at least 10 times for each of the test molecules, which yielded standard deviations (Table I) that ranged from 0.1 to 0.2 Å. For modeling the solvent, we used the GAUSSIAN09 program's default implementation of the integral equation formulation of a polarizable continuum model. The

TABLE I. Estimated $p\mathrm{K}_{\mathrm{int}}$ values used (see the text). Corresponding model $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values for individual residues are given in [122] for $\alpha_*$ equal to the top pattern at $p$H 7.1.

| Residue | Abbreviation | $r_0$±s.d.[a] (Å) | $p\mathrm{K}_{\mathrm{H_2O}}$ | $\Delta p\mathrm{K}$ | $p\mathrm{K}_{\mathrm{int},\varepsilon=3}$ | No. |
|---|---|---|---|---|---|---|
| arginine | Arg(R) | 3.68 ± 0.2 | 12.48 | −1.18 | 11.30 ± 0.2 | 20 |
| aspartate | Asp(D) | 3.31 ± 0.15 | 3.86 | +1.55 | 5.41 ∓ 0.2 | 13 |
| cysteine | Cys(C) | 3.35 ± 0.2 | 10.50[b] | +1.50 | 12.00 ∓ 0.2 | 3 |
| glutamate | Glu(E) | 3.31 ± 0.15 | 4.25 | +1.55 | 5.80 ∓ 0.2 | 9 |
| *N*-glycine | Gly(G) | 2.65 ± 0.2 | 7.60 | −2.72 | 4.88 ± 0.5 | 1 |
| histidine | His(H) | 3.65 ± 0.2 | 6.00 | −1.20 | 4.80 ± 0.2[c] | 5 |
| histidine | H14 | | | | 7.05[d] | 1 |
| histidine | H53 | | | | 6.04[d] | 1 |
| histidine | H84 | | | | 6.91[d] | 1 |
| histidine | H117 | | | | 6.37[d] | 1 |
| histidine | H122 | | | | 6.32[d] | 1 |
| lysine | Lys(K) | 2.65 ± 0.2 | 10.70 | −2.72 | 7.98 ± 0.5 | 2 |
| C-tyrosine | Tyr(Y) | 3.31 ± 0.15 | 3.40 | +1.55 | 4.95 ∓ 0.2 | 1 |

[a] Ions used in GAUSSIAN09 calculations with $H_2O$ solvent were methylguanidinium(Arg); acetate[Asp, Glu, Y174(carboxyl)]; ammonium[Lys, G1(amino)]; mean of imidazolium, 4-methyl imidazolium(His); and methanethiolate(Cys).

[b] See the text.

[c] Value not used; see the text.

[d] From PROPKA 3.1 [115–118]. Note that these values are not intended as $p\mathrm{K}_{\mathrm{int}}$ values; see Sec. II F for discussion.

needed $p\mathrm{K}_{\mathrm{H_2O}}$ values were estimated with use of the tables given in the work of Dawson *et al.* [98], Ellenbogen [99], and Serjeant and Dempsey [100]. Because this work is spurred by our interest in building models for $\gamma$B-$\gamma$B interactions in the range $4 < p\mathrm{H} < 8$, we did not include the titration of tyrosine side chains, which typically occurs in the range $10 < p\mathrm{H} < 10.3$ [110].

The resulting $p\mathrm{K}_{\mathrm{int}}$ values and uncertainties are listed in Table I. The values calculated for the charged sites just inside the low-dielectric sphere are designated as $p\mathrm{K}_{\mathrm{int},\varepsilon=3}$. Here we are anticipating the fact that, as explained below, the grand-canonical distribution model was used to predict titration curves as functions of $\varepsilon$, which were then compared with experiment to settle on an assumed, continuum model internal static dielectric coefficient value $\varepsilon = 3$.

However, when carrying out this process, we found that there was a discrepancy between the modeled titration curve and the data, displayed in Fig. 4(a) below, which suggested that our modeled values for the histidine $p$K were lower than would be compatible with the titration curve [73] and the measured isoelectric point of bovine $\gamma$B-crystallin, $p\mathrm{H} = 7.8$ [74,75]. Therefore, as input to the grand-canonical simulations we instead tried using the PROPKA (version 3.1) web server estimates [115–118] to replace the initially estimated histidine $p\mathrm{K}_{\mathrm{int}}$ values for $\gamma$B, again with use of the PDB entry 1AMM, while leaving all the other $p\mathrm{K}_{\mathrm{int}}$ values to remain estimated by the $\mathbf{D}\cdot\mathbf{E}$ integral method described above. The resulting PROPKA histidine $p$K estimates are listed in Table I. The comparison of the modeled titration curve with the data, for various assumptions about the inner dielectric coefficient and the histidine $p$K values, is described and shown in Sec. II F below, in connection with Fig. 4. As the authors emphasize [115–118], PROPKA uses a phenomenological approach to achieve speed and scope in estimating $p$K values for a large variety of proteins. It incorporates factors that the present approach does not, including hydrogen bonding and varying degrees of penetration of residues into the interior. For histidines PROPKA starts from a higher model $p$K of 6.5 than the 6.0 initially used here and it assigns smaller $p$K reductions to H14, H84, H117, and H122 than the Table I $\Delta p$K of −1.20, due to their varying degrees of penetration. PROPKA also predicts that hydrogen bonds raise the $p$K values of H14, H53, and H84, by from 0.6 to almost 0.9 $p$K units.

The Table I cysteine $p\mathrm{K}_{\mathrm{H_2O}}$, 10.5, chosen to be near the $p$K values for methanethiol (10.33) and ethanethiol (10.5 and 10.61) [100], leads to $p\mathrm{K}_{\mathrm{int}}$ values of 12, well above the 9–9.5 typical of protein cysteines [110]; indeed, many proteins exhibit cysteine $p\mathrm{K}_{1/2}$ values much less than 9, due primarily to hydrogen bonding [119] and somewhat to nearby amide dipole potentials [47,119]. We did not alter the present approach given our focus on $4 < p\mathrm{H} < 8$, although better modeling of cysteine $p$K values is of interest, given the importance of cysteine oxidation for gamma crystallins and other lens proteins [54–56,120,121]. In particular, the present model does not attempt to model hydrogen bonding of C18 with both C78 and S20, predicted by PROPKA to lower the C18 $p$K to 6.88. We included only C15, C18, and C22 in our model, which appear less buried than C32, C41, C78, and C109. However, due to the high model cysteine $p\mathrm{K}_{\mathrm{int}}$, C15, C18, and C22 were charged only rarely in simulations, at higher $p$H values, as tabulated in the Supplemental Material [122].

We will find it instructive to study the ability of "effective" $p$K values, $p\mathrm{K}_{\mathrm{eff},\alpha_*}$, defined below, to model the probability distributions of the protonation patterns. These effective $p$K values are closely related to those used in Refs. [27–29], among others; briefly, the difference is that here we study the $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ with respect to particular choices $\alpha_*$ of on-or-off charge patterns $\mathbf{Q}_b + \mathbf{O}_{\alpha_*}$, as contrasted with patterns of average residue charge values at a given $p$H, $\mathbf{Q}_b + \langle\mathbf{O}_{\alpha_*}\rangle$ in the present notation, as analyzed, for example, in Ref. [29].

In the present notation, the $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ are expressed as follows [27]. For any particular protonation pattern $\alpha$ on the protein, the numerator of Eq. (5) can be written

$$\mathcal{Q}P_\alpha \equiv \tilde{\mathcal{Q}}_\alpha = 10^{(p\mathbf{K}-p\mathbf{H})\cdot\mathbf{O}_\alpha} e^{-\mathbf{q}_\alpha^\top W \mathbf{q}_\alpha/2}, \quad \mathbf{q}_\alpha = \mathbf{Q}_b + \mathbf{O}_\alpha, \tag{7}$$

in which $\mathbf{Q}_b$ is the vector of bare charges of the titratable residues; $p\mathbf{K}$, $p\mathbf{H}$, and $\mathbf{q}_\alpha$ also denote vectors; and the symmetric work-of-charging $W$ is defined above Eq. (4). The idea is now to use a chosen configuration $\mathbf{O}_{\alpha_*}$, which could be, for example, the most probable configuration at a certain $p$H, as a reference configuration; the algebraic development given here also applies for the reference configuration choice $\langle\mathbf{O}_\alpha\rangle$, as used in Refs. [27–29]. The probabilities of other configurations can now be expressed in terms of how much their occupancy vectors differ from that of the reference configuration. For configuration $\alpha$, the occupancy vector is $\mathbf{O}_\alpha = \mathbf{O}_{\alpha_*} + (\mathbf{O}_\alpha - \mathbf{O}_{\alpha_*}) = \mathbf{O}_{\alpha_*} + \delta\mathbf{O}_\alpha$, where $\delta\mathbf{O}_\alpha = \mathbf{O}_\alpha - \mathbf{O}_{\alpha_*}$. Letting $\mathbf{q}_{\alpha_*} = \mathbf{Q}_b + \mathbf{O}_{\alpha_*}$ and $\mathbf{q}_\alpha = \mathbf{q}_{\alpha_*} + \delta\mathbf{O}_\alpha$, one finds

$$\begin{aligned}\tilde{\mathcal{Q}}_\alpha &= 10^{(p\mathbf{K}-p\mathbf{H})\cdot(\mathbf{O}_{\alpha*}+\delta\mathbf{O}_\alpha)} e^{-(\mathbf{q}_{\alpha*}+\delta\mathbf{O}_\alpha)^\top W(\mathbf{q}_{\alpha*}+\delta\mathbf{O}_\alpha)/2}\\ &= 10^{(p\mathbf{K}-p\mathbf{H})\cdot\mathbf{O}_{\alpha*}} e^{-\mathbf{q}_{\alpha*}^\top W\mathbf{q}_{\alpha*}/2} 10^{(p\mathbf{K}-p\mathbf{H})\cdot\delta\mathbf{O}_\alpha}\\ &\quad\times e^{-(2\delta\mathbf{O}_\alpha^\top W\mathbf{q}_{\alpha*}+\delta\mathbf{O}_\alpha^\top W\delta\mathbf{O}_\alpha)/2}.\end{aligned}$$

The first two multiplicative factors in the above expression are common to all $\tilde{\mathcal{Q}}_\alpha$. In the expression for the probabilities $P_\alpha$, these factors cancel with the same common factors in the denominator, $\mathcal{Q}$. Therefore, we have $P_\alpha = \tilde{\mathcal{Q}}_{\alpha_*}(\alpha)/\mathcal{Q}_{\alpha_*}$, in which we define $\tilde{\mathcal{Q}}_{\alpha_*}(\alpha)$ and $\mathcal{Q}_{\alpha_*}$ via

$$\begin{aligned}Q_\alpha &\equiv 10^{(p\mathbf{K}-p\mathbf{H})\cdot\mathbf{O}_{\alpha*}} e^{-\mathbf{q}_{\alpha*}^\top W\mathbf{q}_{\alpha*}/2}\tilde{\mathcal{Q}}_{\alpha_*}(\alpha),\\ \mathcal{Q} &\equiv 10^{(p\mathbf{K}-p\mathbf{H})\cdot\mathbf{O}_{\alpha*}} e^{-\mathbf{q}_{\alpha*}^\top W\mathbf{q}_{\alpha*}/2}\mathcal{Q}_{\alpha_*}.\end{aligned}$$

With these definitions,

$$\begin{aligned}\tilde{\mathcal{Q}}_{\alpha_*}(\alpha) &= 10^{(p\mathbf{K}-p\mathbf{H})\cdot\delta\mathbf{O}_\alpha} e^{-(2\delta\mathbf{O}_\alpha^\top W\mathbf{q}_{\alpha*}+\delta\mathbf{O}_\alpha^\top W\delta\mathbf{O}_\alpha)/2}\\ &= 10^{(p\mathbf{K}-p\mathbf{H})\cdot\delta\mathbf{O}_\alpha} e^{-(W\mathbf{q}_{\alpha*})\cdot\delta\mathbf{O}_\alpha} e^{-\delta\mathbf{O}_\alpha^\top W\delta\mathbf{O}_\alpha/2}\\ &= 10^{(p\mathbf{K}-p\mathbf{H})\cdot\delta\mathbf{O}_\alpha} 10^{-(W\mathbf{q}_{\alpha*})\cdot\delta\mathbf{O}_\alpha/\ln 10} e^{-\delta\mathbf{O}_\alpha^\top W\delta\mathbf{O}_\alpha/2}\\ &= 10^{[(p\mathbf{K}-W\mathbf{q}_{\alpha*}/\ln 10)-p\mathbf{H}]\cdot\delta\mathbf{O}_\alpha} e^{-\delta\mathbf{O}_\alpha^\top W\delta\mathbf{O}_\alpha/2}.\end{aligned} \tag{8}$$

With $\delta\mathbf{q}_\alpha = \mathbf{q}_\alpha - \mathbf{q}_{\alpha_*} = \delta\mathbf{O}_\alpha$, Eq. (8) can be written as

$$\tilde{\mathcal{Q}}_{\alpha_*}(\alpha) = 10^{[(p\mathbf{K}-W\mathbf{q}_{\alpha*}/\ln 10)-p\mathbf{H}]\cdot\delta\mathbf{O}_\alpha} e^{-\delta\mathbf{q}_\alpha^\top W\delta\mathbf{q}_\alpha/2}. \tag{9}$$

From a comparison of Eqs. (7) and (9), it is natural to define a vector $p\mathbf{K}_{\mathrm{eff},\alpha_*}$ of $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values by

$$p\mathbf{K}_{\mathrm{eff},\alpha_*} = p\mathbf{K} - W\mathbf{q}_{\alpha_*}/\ln 10, \tag{10}$$

so that

$$\tilde{\mathcal{Q}}_{\alpha_*}(\alpha) = 10^{(p\mathbf{K}_{\mathrm{eff},\alpha*}-\,p\mathbf{H})\cdot\delta\mathbf{q}_\alpha} e^{-\delta\mathbf{q}_\alpha^\top W\delta\mathbf{q}_\alpha/2}, \tag{11}$$

$$P_\alpha = \frac{\tilde{\mathcal{Q}}_{\alpha_*}(\alpha)}{\mathcal{Q}_{\alpha_*}} \quad \text{with } \mathcal{Q}_{\alpha_*} = \sum_\alpha \tilde{\mathcal{Q}}_{\alpha_*}(\alpha). \tag{12}$$

Equations (10)–(12) correspond to the combination of Eqs. (2)–(4) of Ref. [27], as expressed there in terms of a particular constellation of charges (here symbolized by the vector $\mathbf{q}_{\alpha_*}$). As indicated above, except by its use of a particular on-off pattern $\alpha_*$, Eq. (10) is also related to Eqs. (1a), (1b), and (16) in Ref. [29], where the average $p$H-dependent approach introduced in Ref. [27] is expressed and its mean-field-approximation nature is elucidated. In this connection we note that the use of $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ below, to study its capability to approximate probability distributions of protonation patterns, has a different focus than the study of the reduced-site approximation also introduced in Ref. [29]; that approximation becomes better as the criteria to regard less-labile sites as fixed become progressively more strict. Reference [29] demonstrates that the reduced-site approximation is more effective than the mean-field approach for representing the average occupancy states of particular sites, while typically more efficient computationally than using the exact expressions.

Equation (10) expresses the fact that for configurations that are similar to the chosen configuration $\alpha_*$, the effective $p$K values are typically biased by charges $\mathbf{q}_{\alpha_*}$ on neighboring sites. These charges, in turn, produce voltages that bias the occupancy of a given site. Thus, at a given $p$H, one expects that the site occupancies can be fairly well described by $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values for a well-chosen $\alpha_*$, say, the most common configuration. The extent to which this is *not* the case is clearly a function of the quantities $e^{-\delta\mathbf{q}_\alpha^\top W\delta\mathbf{q}_\alpha/2}$, according to Eqs. (11) and (12). We will find below that a given set of $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values accurately represent only *part* of the probability distribution of the protonation patterns at a given $p$H, precisely because of this latter factor.

### E. Calculation of the potential and work-of-charging matrix

The numerical methods we used to calculate the potential are those described previously [10]. We used grid sizes from 0.3 to 0.6 Å, the domain was $100 \times 100 \times 120$ Å$^3$, and the protein was placed at its center. We used the Neumann boundary condition that the normal component of the field is zero there. While we expect that a more accurate boundary condition would be that a linear combination of the normal field component and the potential would be zero, for the Debye lengths investigated here, the zero-field condition suffices. To calculate the $i$th row of the work-of-charging matrix $W$, a charge is placed at site $i$; the potential at site $j$ then gives the entry $w_{ij}$. Each such work-of-charging matrix was symmetric, providing an important check on the calculation.

We note that there are also self-energies associated with the interaction of each charge with its counterion cloud. In principle, this factor also changes the effective $p$K of a site, above and beyond the fact that the site is near a dielectric boundary. We calculated the magnitudes of these effects from our numerical solutions of Eq. (1), by evaluating the potential at a given charged site produced by the nearby net charge within its surrounding, screening ionic atmosphere. The magnitudes we calculated for this effect were very uniform and would produce changes in the given $p$K on the order of only ±0.1 $p$K units, which we regard as insignificant compared with the uncertainties in the modeled $p\mathrm{K}_{\mathrm{int}}$ values themselves. Accordingly, we simply set the diagonal entries of the work-of-charging matrices to 0 for further calculations.

Figure 3 illustrates a work-of-charging matrix calculated in this fashion. To find a permutation of the residue order that would yield the approximately block-diagonal forms shown

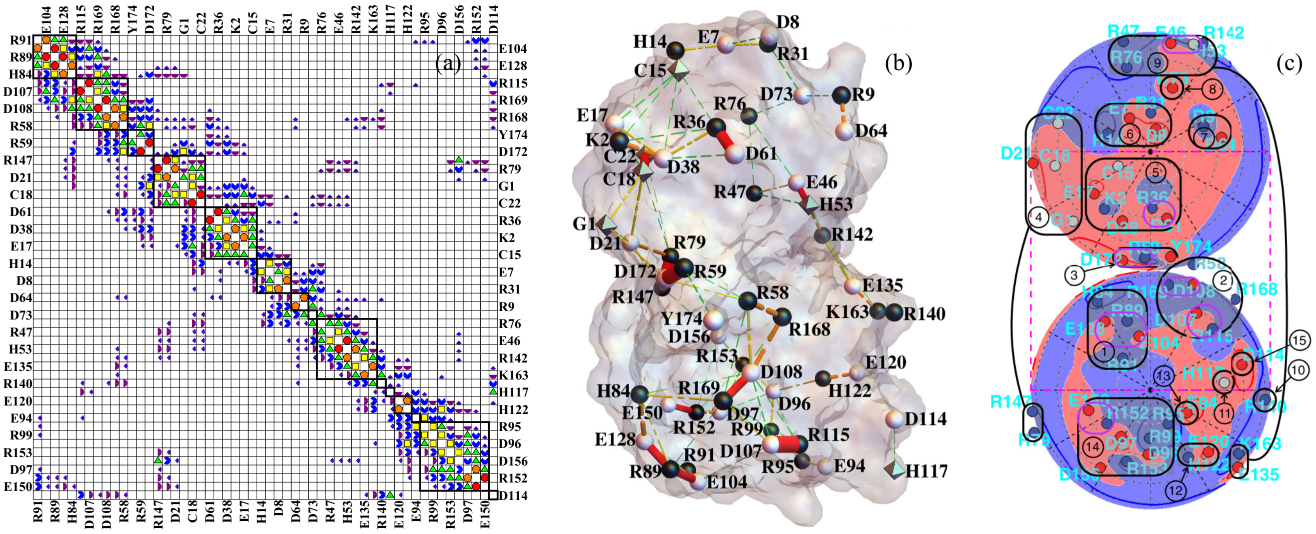

FIG. 3. (a) Approximate block-diagonal form of the dimensionless work-of-charging matrix $W$ for interior dielectric coefficient 3 and Debye length 6.0 Å. The $W_{ij}$ magnitude categories are as follows: white $< 0.05 \leqslant$ dark purple quarter circles $< 0.1 \leqslant$ purple half circles $< 0.2 \leqslant$ blue 3/4 circles $< 0.4 \leqslant$ green triangles $< 0.8 \leqslant$ yellow squares $< 1.6 \leqslant$ orange pentagons $< 3.2 \leqslant$ red circles. The matrix includes all 54 titratable residues used in the present model, which as noted in the text omits the tyrosine residues and four of the cysteine residues; the entire protein contains 174 residues. Designations for the 54 residues considered alternate between left and right (and top and bottom) margins. (b) Cylinders with radii proportional to $W_{ij}$ mapped onto the PDB 1AMM structure of $\gamma$B-crystallin. The $W_{ij}$ magnitude categories are as follows: $0.4 \leqslant$ green cylinders with three gaps $< 0.8 \leqslant$ yellow cylinders with two gaps $< 1.6 \leqslant$ orange cylinders with one gap $< 3.2 \leqslant$ red cylinders. (c) Lambert projections with potential and charges indicated as in Fig. 1(c). Groups of titratable sites participating in approximate blocks of $W$ are circled in black and numbered in (c) and indicated by black squares in (a). Group numbering corresponds to the order of sites in $W$ in Fig. 3(a), from top to bottom, and group numbers are those to which Figs. 7 and 8 refer. In (b) and (c) the protonation configuration is that modeled to be the most common one at $p$H 7.1.

in Fig. 3, we used simulated annealing, with an objective function that was linearly proportional to the distance of (the symmetric) work-of-charging entries from either the diagonal or the upper right or lower left corners. On repeated runs, this yielded a robust grouping of sites. While we grouped the sites in this fashion in order to identify patches of residues predicted by the model to be more highly correlated, we left all entries intact for computing the partition function. That is, this grouping does not represent a block-diagonal approximation method, an avenue that has been pursued by a number of investigators (see Ref. [123] and references therein).

Figure 3(a) displays an approximate block-diagonal form of the work-of-charging matrix $W$, for the adopted inner dielectric value $\varepsilon_{\text{in}} = 3.0$. The symbol code for $W_{ij}$ magnitude categories is given in the caption, ranging from white for entries less than $0.05k_BT/e$ to red circles for entries greater than or equal to $3.2k_BT/e$. The prominent entries adjacent to the main diagonal show a high degree of charge pairing, long noted to occur for $\gamma$-crystallins [124]. The 0.05 lower cutoff is close to the value below which we observed very little change in the order of probabilities of the protonation patterns, if smaller entries were ignored [see Fig. 10(a)]. Residue identities are indicated on the borders of Fig. 3(a). A perspective view of the work-of-charging matrix of Fig. 3(a) is given in Fig. 2 of the Supplemental Material [122]. Figure 3(b) displays the work-of-charging entries in the form of line segments that link the titratable groups on the protein, using the same symbol code as in Fig. 3(a). Figure 3(c) shows labeled sets of titratable sites, circled in black, that participate in approximate blocks of $W$, with use of the same projection as in Fig. 1(c). The corresponding blocks are outlined by the thick black squares in Fig. 3(a). In addition, prominent charge pairs are circled in purple (lighter) in Fig. 3(c). Tables of the work-of-charging matrices we calculated for Debye lengths 6, 12, and 20 Å are given in Figs. 6–11 of the Supplemental Material [122].

### F. Calculation of the grand-canonical distribution function and the protonation pattern probabilities

We performed Metropolis Monte Carlo simulations that included all 54 sites of the present model to determine the grand-canonical partition function (GCPF) and the associated statistics of the distributions of protons on the protein. Protonation pattern statistics were studied using Monte Carlo runs of $10^8$ iterations. We determined GCPF vs $p$H in 0.1 $p$H increments by finding top protonation configuration probabilities in $10^6$ iteration runs and using Eq. (5) with that configuration's $\Delta G_\alpha$. While the results given here were calculated from the simulations, it is convenient to note that in the Monte Carlo simulations, many of the residues, primarily the arginines with the highest $p\text{K}_{\text{eff},\alpha_*}$ values, never changed their occupation states at some of the $p$H values in the range of primary interest here, 4–8, or did so very few times, even in $10^8$ iterations. A table of the number of times each residue switched protonation state, as a function of $p$H, and a table that includes individual $p\text{K}_{\text{eff},\alpha_*}$ values appear as Figs. 4 and 5, respectively, in the Supplemental Material [122]. Therefore, to speed calculations, it can be convenient to omit such residues from calculation of the partition function, as was done in the reduced sites approximation of Ref. [29], and with fewer

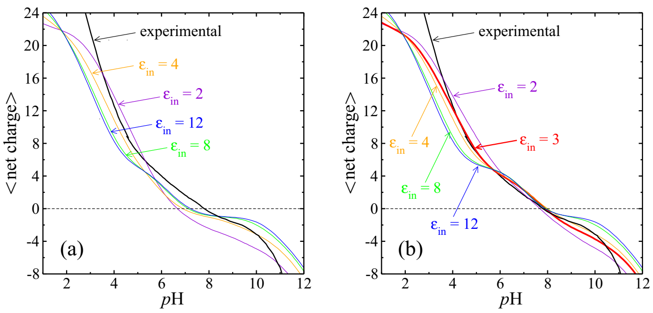

FIG. 4. Selection of interior dielectric coefficient and $p$K values through comparison of modeled titration curves with experiment. The experimental data [73] are shown by the labeled curve. The work-of-charging matrix $W$ was calculated as described in the text, for different choices of $\varepsilon_{\text{in}}$. (a) Calculated titration curves when all $p\text{K}_{\text{int},\varepsilon_{\text{in}}}$ values were calculated according to the $\mathbf{D}(\mathbf{r}) \cdot \mathbf{E}(\mathbf{r})$ integral method described in Sec. II D, for the same $\varepsilon_{\text{in}}$ values used for solution of Eq. (1) to yield the matrix $W$. (b) Calculated titration curves when all but the histidine $p\text{K}_{\text{int},\varepsilon_{\text{in}}}$ values were estimated with the integral method, as functions of $\varepsilon_{\text{in}}$, while PROPKA 3.1 values were used for histidine (Table I) $p\text{K}_{\text{int}}$ values (see the text). In (b) the red (bold) $\varepsilon_{\text{in}} = 3$ curve is that of the model adopted for further study of the probability distributions.

titratable sites, about 25 or 30, the model partition function $\mathcal{Q}$ can be evaluated exactly. By either method, once $\mathcal{Q}$ is known, the probability of protonation pattern $\alpha$ is then given by Eq. (5). Likewise, the average number of protons $\langle n \rangle$ on a protein can be found from

$$\langle n \rangle = \frac{\zeta}{\mathcal{Q}} \frac{\partial \mathcal{Q}}{\partial \zeta} = \frac{\partial \ln \mathcal{Q}}{\partial \ln \zeta}, \tag{13}$$

in which $\zeta = 10^{-p\text{H}}$.

Titration curves calculated in this manner are shown in Fig. 4. The experimental data [73] are shown by the black curve in each panel. These data were obtained with use of an aqueous 100 mM potassium chloride solvent, corresponding to a Debye length of 9.6Å, the value we therefore used in the 54 solutions of Eq. (1) for each choice of $\varepsilon_{\text{in}}$, to generate the matrices $W$ needed for the comparisons shown in Fig. 4. Figure 4(a) shows the calculated titration curves when the $p\text{K}_{\text{int},\varepsilon_{\text{in}}}$ values were calculated according to the $\mathbf{D} \cdot \mathbf{E}$ integral method described above. Note that in this case *both* the work-of-charging matrix $W$ resulting from application of Eq. (1) *and* the $p\text{K}_{\text{int},\varepsilon_{\text{in}}}$ values resulting from the first two lines of Eq. (6) are functions of the interior dielectric coefficient value $\varepsilon_{\text{in}}$ and were calculated as input to the GCPF simulations for the values $\varepsilon_{\text{in}} = 2, 4, 8$, and 12. Therefore, the calculated $p\text{K}_{\text{int},\varepsilon_{\text{in}}}$ values relevant to the curves in Fig. 4(a) are not those listed in Table I for $\varepsilon_{\text{in}} = 3$. In Fig. 4(a) each test model titration curve predicts a lower isoelectric point ($p$I) than that observed experimentally for $\gamma$B-crystallin, $p\text{I} = 7.8$ for the native protein [74,75], though in Ref. [74] a minor component was also observed at a lower $p$I of 7.3, a component that was sensitive to the presence of reducing agents [74]. Also, note that bovine $\gamma$B-crystallin was termed $\gamma$-II at the time of publication of Refs. [74,75].

In addition to the purpose of studying the probability distributions of the protonation patterns, we have a goal of modeling small-angle neutron scattering data from $\gamma$B-crystallin solutions in the $p$H range between 4.5 and 7.1, and as a preliminary step want to create a charge-regulation model in a $p$H range that spans these values and reproduces the observed isoelectric point. Therefore, as described above, we used the PROPKA estimates for the needed histidine $p\text{K}_{\text{int}}$ values, while continuing to use the $\mathbf{D} \cdot \mathbf{E}$ integral procedure for the other residues. Figure 4(b) shows the resulting model titration curves. Although there is clearly a range of $\varepsilon_{\text{in}}$ values that could be used and there is room for improvement, the highlighted

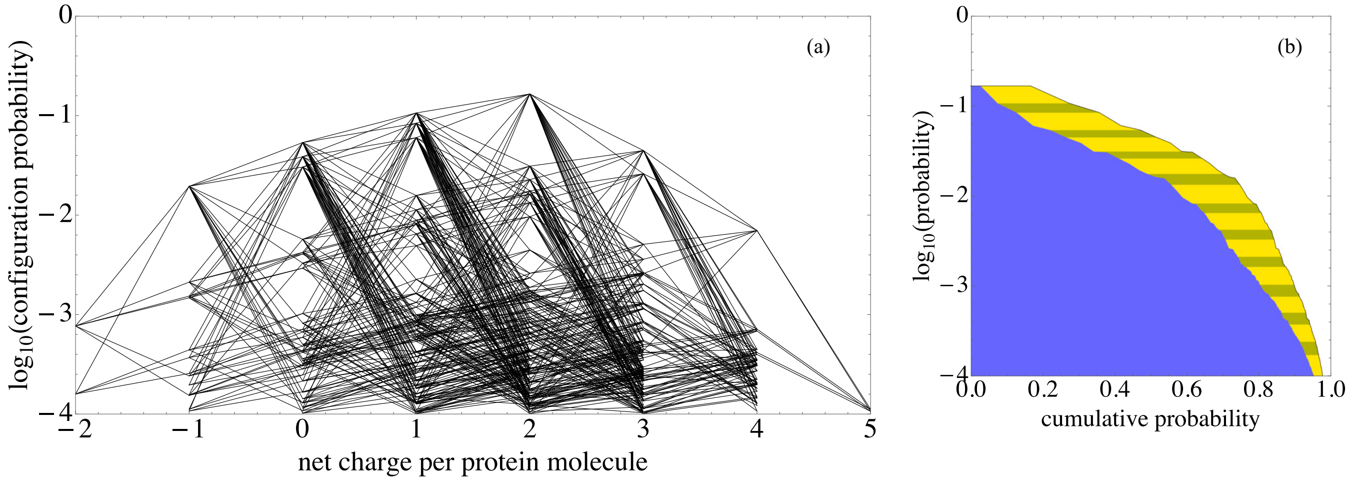

FIG. 5. Protonation patterns that have opposite net protein charge readily occur at $p$H 7.1. In both panels, $\log_{10} P$ is plotted vertically for the most prominent $p$H 7.1 configurations that together account for over 97% of the configuration probability. (a) The net protein charge of each configuration is plotted horizontally. Line segments join configurations that can be transformed into one another with a single-residue protonation switch. (b) (i) The horizontal coordinate of the yellow-striped–clear boundary is the sum of the probabilities of that configuration and more common ones, that is, their cumulative probability. (ii) The horizontal coordinate of the blue–yellow-striped boundary is the square of the same cumulative probability. The blue–yellow-striped boundary estimates the fraction of pairs of neighboring proteins, both molecules of which have one of the configurations down to a given $\log_{10} P$ level; this estimate neglects biasing of pattern probabilities due to protein proximity.

red curve, with $W$ and nonhistidine $p$K values generated using $\varepsilon_{\text{in}} = 3.0$, provides a relatively good match to the experimental titration curve in the range $4 < p\text{H} < 8$ of particular interest and we took it to be sufficient for studying the general nature of the protonation pattern probability distributions. We did so despite the fact that the PROPKA estimates include a model of charge-charge interactions [117] and therefore are not intended to be intrinsic $p\text{K}_{\text{int}}$ values as they are used here.

We anticipate that as NMR assignment and titration data become available for $\gamma$B-crystallin, it will become possible to test and refine the present model in much more detail. Accordingly, we postponed detailed study of using different histidine $p\text{K}_{\text{H}_2\text{O}}$ values, which are expected to depend on their tautomeric states [125], in addition to the possible hydrogen bonding, dipolar potential, and other effects mentioned above.

The value $\varepsilon_{\text{in}} = 3.0$ of the model we use here is compatible with calculations of continuum-model static dielectric coefficients of 2–4 for *interior* regions of many proteins and with measurements of dry protein powders [90,126–129]. The quoted range is approximate, depending on the protein and the method of calculation, and represents an ongoing area of investigation, as noted above [92]. Recent analyses of NMR chemical shifts within proteins, in particular their dependence on modeled local electric fields, found that values of $\varepsilon_{\text{in}}$ near 3 gave the best matches to data [130,131].

## III. PROBABILITY DISTRIBUTIONS OF PROTONATION PATTERNS

### A. Features of the distributions at constant $p$H

In this section we address the following questions. How broad are the distributions at a given $p$H? How different are these distributions from the multinomial distribution that would occur if the off-diagonal work-of-charging entries were all zero? What simple approximations provide good quantitative agreement with the exact model probabilities? How different are the patterns of surface voltage that correspond to probable protonation patterns? In order to study these questions, in Figs. 5, 7, and 8 we plot the base-10 logarithm of the modeled probability of each protonation pattern vertically vs its net charge. In each figure, the line segments join two configurations that differ by a single switch in proton occupancy. We call such configurations adjacent. In addition to giving a visual picture of the protonation pattern probabilities and the possible single-step transitions between them, these and related diagrams can help to study how pattern probabilities are distributed with respect to factors that can affect protein-protein interactions, here net charge.

Figure 5 shows that $\gamma$B protonation patterns that have both positive and negative net protein charge readily occur at $p$H 7.1. Figure 5(a) shows the most prominent $p$H 7.1 configurations that together account for 97% of the configuration probability.

In Fig. 5(b) the cumulative probability down to a given level is plotted horizontally as the curve on the far right, the boundary of the striped yellow region. This curve, taken together with Fig. 5(a), shows that each of the topmost 70% of the configurations has a non-negative net charge. However, below that level quite a few $p$H 7.1 configurations have net negative charge. Because oppositely charged proteins are more likely to exhibit attractive interactions, Fig. 5 gives rise to the interesting possibility that at high concentrations, where proteins have many near neighbors, the probability distributions of net charge may even become bimodal. In this work we do not analyze the biasing of the distributions because of protein proximity.

The *square* of the cumulative probability is filled in blue (dark) in Fig. 5(b). It provides an estimate of the fraction of *pairs* of neighboring proteins, both molecules of which have a configuration with a probability above a given level. This estimate again neglects biasing of probabilities due to protein proximity. The blue-yellow boundary suggests that close pairs of proteins, both of which have net non-negative charge, will account for only the top half of neighboring protein pairs. Also, to account for about 80% of the configuration pair types, configurations that range down to those that occur only one one-thousandth of the time must be included. Thus the blue-yellow boundary gives a rough guide to how many configurations to include in a model of electrostatic interactions for this protein.

Figure 6 compares the voltage patterns around the 12 most probable proton configurations at $p\text{H} = 7.1$ and Debye length 6 Å. Residues that have gained or lost protons, with respect to the next more common configuration, are shown by blue (darker) and red (lighter) arrows, respectively. At this $p$H, histidine protonation switches are modeled to account for the first 20 patterns. It is very interesting that as a consequence of the majority of these switches, the connectivity of the positive [blue (darker)] and negative [red (lighter)] potential regions on the projection spheres also changes, much like straits and isthmuses in continental drift. Thus one might expect that in the presence of neighboring proteins that also have charged patches, the ease of reorientation of each protein could depend on voltage channels that open and close, as each of their protonation configurations changes. The similarity of many of the voltage patterns that result from different protonation patterns, illustrated in Fig. 6, suggests that larger classes of such pairs may be sufficient for creating accurate models of the relevant pair potentials. Thus a very interesting question is how best to construct a good coarse-grained level of detail in the protonation pattern distributions in order to model protein interactions accurately. In the present work we do not focus on the protein interaction consequences of the patterns shown in Fig. 6.

We now study the origins of the switching pattern shown in Fig. 5(a) in more detail, in a residue-by-residue manner. Each line segment in Fig. 5(a) can be identified with the particular residue that gained or lost a proton. The probabilities of protonation patterns reflect both the affinity of each residue for protons and the correlations between sites that are strongly affected by their mutual electrostatic interaction.

The quantitative consequences are illustrated in Fig. 7. If two residues are uncorrelated, as are H122 and H84, the change in the pattern probability when one of them switches protonation state will not depend on the state of the other. Because their occupation probabilities are essentially independent, when H122 changes its charge, the logarithm of the pattern probability will change by a given amount that does not depend on whether H84 is protonated. Thus, the slope of the line segment that links H122-adjacent patterns (purple solid line) will not depend on the state of H84 and vice versa. In

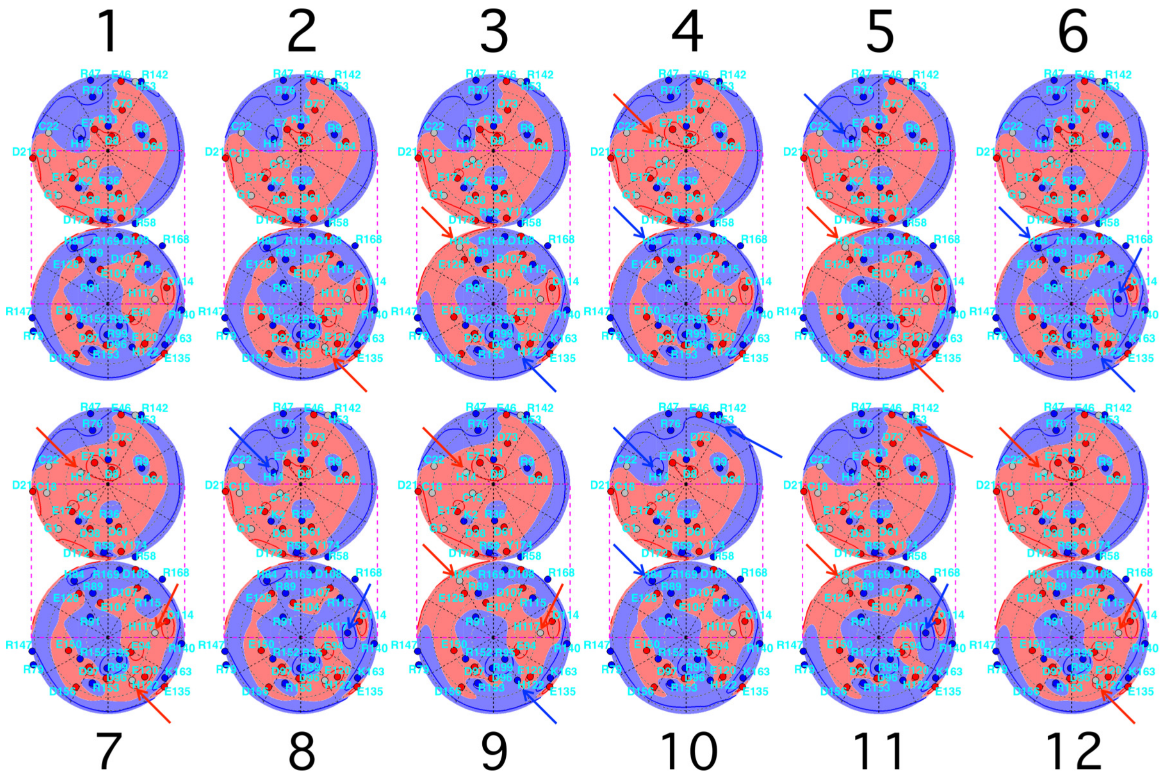

FIG. 6. Lambert projections, with potentials and charges indicated as in Fig. 1(c), for the 12 most probable configurations at $p\mathrm{H} = 7.1$ and Debye length 6 Å, in order of probability (see Table II). Residues that have gained or lost a proton, with respect to the more common configuration that is adjacent in order, are shown by blue (darker) and red (lighter) arrows, respectively. For many protonation switches, positive [blue (darker)] and negative [red (lighter)] voltage regions change connectivity.

contrast, if two sites are strongly correlated, their protonation probabilities are no longer independent and the corresponding slopes that link adjacent configurations will depend on the protonation of the second residue.

Consider Figs. 7(a) and 7(b). In Fig. 7(a), because residue H117 is uncorrelated with residues H122, H84, and H14, protonating H117 simply translates (red short-dashed lines) the line segments for switches of the three other residues. In Fig. 7(b), because E120 is in residue group 12, as is H122 [see the lower right corner of Fig. 3(c)], H122-adjacent pattern probabilities change in different ways that depend on the state of E120.

If the work-of-charging matrix were diagonal, the distribution of protonation patterns would be multinomial and a translation-without-distortion property would hold *exactly* for all line segments in a diagram such as the ones in Figs. 5, 7, and below in Fig. 8. Thus the deviations from congruence of residue-switch polygons in the coordinates (net protein charge $\log_{10} P$) display the degree to which parts of the protonation pattern distribution differ from multinomial. We note that in the present work the choice of net charge on the horizontal axis underlies this polygon translation property, simply because the net charge is here assumed to be solely due to protonation switches. Clearly, if ion absorption played a significant role or if other coordinates were used in place of or in addition to net charge, such as the percentage of the surface that has a positive voltage, a more complex picture would result.

Figure 7(c) focuses on the configurations in the lower, eight-vertex polygon in Fig. 7(b). It illustrates the deviation of the probabilities calculated from the full model from those of the

TABLE II. Occupancies of the residues that switch protonation states in the top ten configurations at $p$H 7.1 (see also Fig. 6). Residues switch in the order H122, H84, H14, H117, and H53, consistent with increasing $|p\mathrm{K}_{\mathrm{eff},\alpha_*} - 7.1|$ (see the text).

| | Configuration rank | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Residue | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| H14 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| H53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| H84 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| H117 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| H122 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| probability | 0.165 | 0.106 | 0.084 | 0.060 | 0.054 | 0.045 | 0.039 | 0.031 | 0.030 | 0.026 |

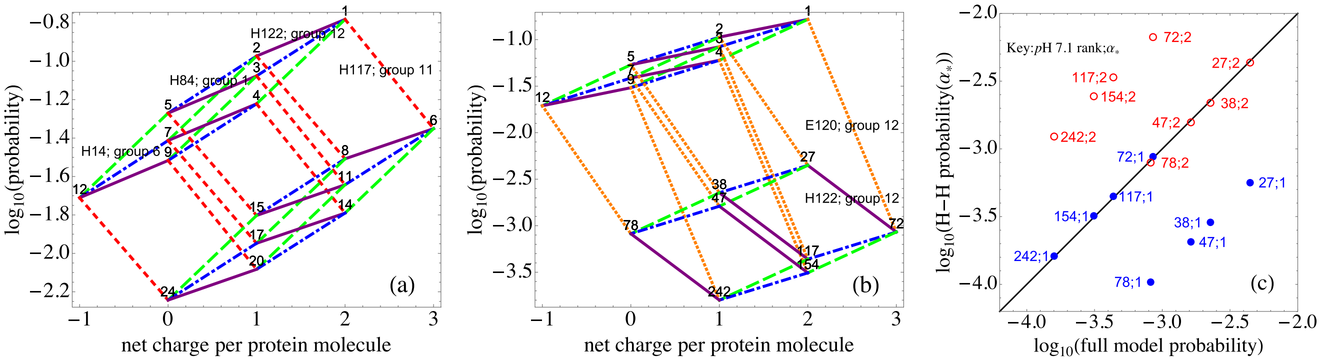

FIG. 7. (a) Because H117 is only weakly linked to other residues [see Fig. 3(c), group 11], when H117 switches charge (red short-dashed lines), the eight-vertex polygon representing the possible switches of H122 (purple solid lines), H84 (blue dash-dotted lines), and H14 (green long-dashed lines) undergoes translation with very little distortion (see the text). (b) Because E120 and H122 interact strongly, when E120 changes from charge −1 to 0 the H122 switching segments markedly change slope, distorting the same polygon. (c) Further analysis of the changes in (b), by comparing the full model probabilities with those of a Henderson-Hasselbalch approach. Agreement would correspond to all points being on the solid diagonal line. (c) Illustration that the E120 switch markedly alters some probabilities from Henderson-Hasselbalch values (see the text).

Henderson-Hasselbalch approximation, in a log-log plot. The blue closed circles compare these two probabilities using the topmost configuration as the reference ($\alpha_* = 1$) for calculating $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values in the Henderson-Hasselbalch approximation. In this case the Henderson-Hasselbalch probability of configuration 72, which is found by a direct single-proton switch from configuration 1, agrees with the probability predicted by the full model, together with the configurations in its attached, translated blue-green polygon, namely, 117, 154, and 242 [see Fig. 7(b)], while the other four configurations (27, 38, 47, and 78) have probabilities that are 10 times those predicted by the Henderson-Hasselbalch model. In contrast, if $\alpha_* = 2$, the open red circles show that the two methods of estimating probability agree for configurations 27, a direct switch from 2, together with 38, 47, and 78, while the other four no longer agree. Because E120 is in the same group as H122, there is no reference state for which all of the pattern probabilities can be computed with use of the Henderson-Hasselbalch approach. Indeed, the factors $e^{-\delta\mathbf{q}_\alpha^\top W \delta\mathbf{q}_\alpha/2}$ in Eqs. (10)–(12), in which the vectors $\delta\mathbf{q}_\alpha$ depend on both $\alpha$ and $\alpha_*$, together with the existence of nonzero off-diagonal elements of $W$, imply that, in general, some full GCPF pattern probabilities will differ from Henderson-Hasselbalch ones, regardless of the choice of $\alpha_*$. Some individual residue protonation probabilities must then also differ from Henderson-Hasselbalch values. This can occur whether or not the residue is charged as it is in $\alpha_*$; this can be shown by expressing individual residue protonation probabilities as sums of the $P_\alpha$ of Eq. (12) over the appropriate patterns $\alpha$, the key point being that each summand can carry a different factor of $e^{-\delta\mathbf{q}_\alpha^\top W \delta\mathbf{q}_\alpha/2}$.

Figure 8 is a larger-scope version of the translating polygons picture. In the present $\gamma$B-crystallin model, at $p$H 7.1 the switching of protonation states of the five histidines accounts for a large fraction of the topmost protonation configurations of

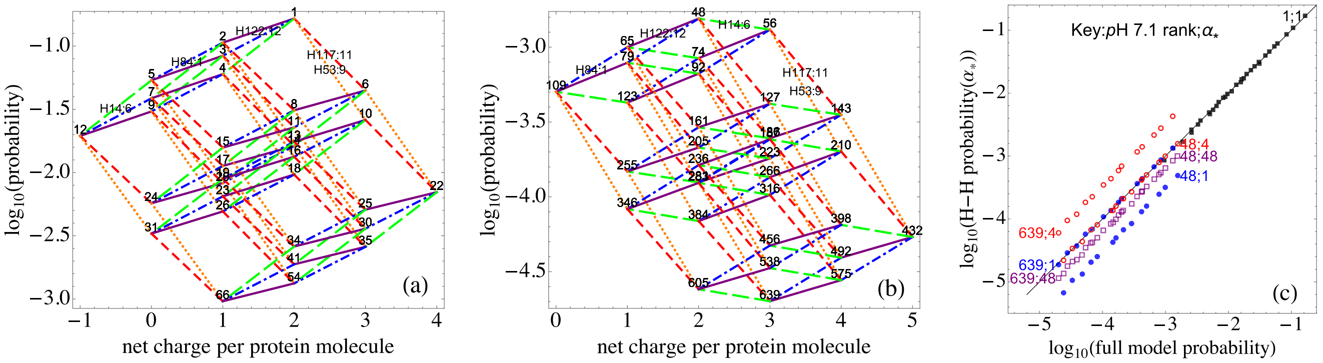

FIG. 8. (a) Most common 32-vertex polygon representing possible switches of all five histidine residues at $p$H 7.1; H53 switches (orange dotted lines) are shown as well as the H122 (purple solid lines), H84 (blue dash-dotted lines), and H14 (green long-dashed lines) depicted in Fig. 7. (b) The 32-vertex polygon of histidine switches that occurs under the condition that E7, a neighbor of H14, has gained a proton to become neutral. The positively charged state of H14, which had been stabilized by a neighboring negative charge, is now less probable than its neutral state and the green long-dashed segments have negative slopes, while the others retain their slopes. Note the change in the vertical scale. (c) Comparison of probabilities from the full model and a Henderson-Hasselbalch approach for the topmost 32-vertex polygon in (a) (black squares), using $\alpha_* = 1$, and the choices $\alpha_* = 1$ (blue closed circles), $\alpha_* = 4$ (red open circles), and $\alpha_* = 48$ (purple open squares) for the polygon in (b). The diagonal solid line is that of agreement between the two methods.

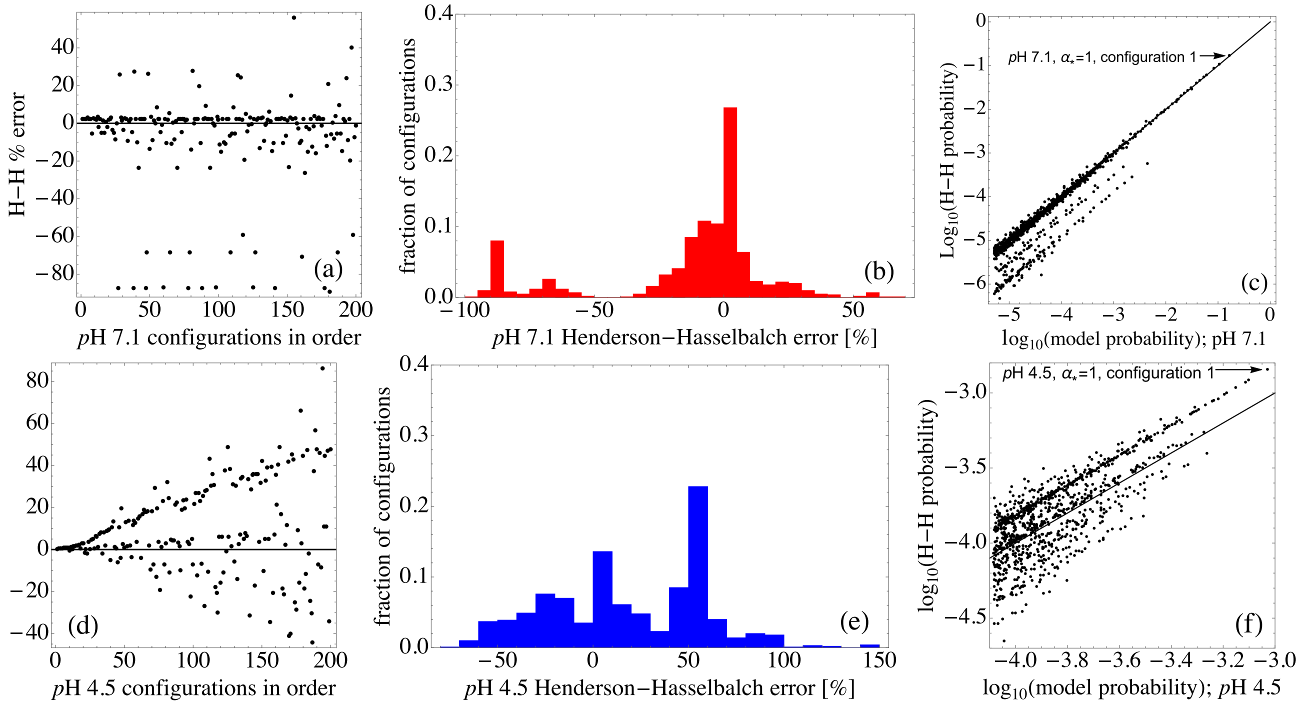

FIG. 9. (a) Percentage deviation of the Henderson-Hasselbalch probabilities of the top-ranked 200 configurations from those of the full model at $p$H 7.1. Panels (a)–(c) all use $\alpha_* = 1$ for $p$H 7.1 to determine $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values for use in calculating Henderson-Hasselbalch probabilities. (b) Histogram of the same deviations, for the top-ranked 1000 configurations at $p$H 7.1. (c) A log-log comparison of the top-ranked 1000 configuration probabilities from the full model with the Henderson-Hasselbalch probabilities. Note the clustering along diagonal lines. (d) Similar percentage deviations of the top-ranked 200 configurations at $p$H 4.5. Panels (d)–(f) all use $\alpha_* = 1$ for $p$H 4.5 to determine $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values for use in calculating Henderson-Hasselbalch probabilities. (e) Histogram of the deviations at $p$H 4.5. (f) A log-log comparison of the top-ranked 1000 configuration probabilities; note the changed scales.

the entire protein. Thus for this $p$H it is interesting to construct sets of 32-vertex polygons, in which the vertices represent *all* of the $2^5$ histidine protonation patterns that occur for a *given* configuration of all the other residues. Figure 8(a) shows the topmost such polygon.

The entire probability distribution of protonation patterns can be represented as the family of all such 32-vertex polygons; each possible pattern belongs to just one such polygon. Figure 8(b) illustrates the distortion of the topmost polygon that results when residue E7, strongly coupled to H14, switches protonation. It is now harder for H14 to become protonated, which is reflected in the fact that the slopes of the green long-dashed segments that represent the H14 protonation switches become smaller; in this case they go from positive to negative. As in Fig. 7, Fig. 8(c) shows that when E7 switches, the Henderson-Hasselbalch approach does not work well for the resulting polygon, even though it does work well for the topmost polygon. The choice $\alpha_* = 48$, suggested by the fact that it is the topmost configuration in Fig. 8(b), produces a linear arrangement that is parallel to but displaced from the line of agreement.

Figure 9 compares full model configuration probabilities with those calculated using $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values, at $p$H 4.5 and 7.1. Figures 9(a), 9(b), 9(d), and 9(e) show that a large number of the configurations have quite different probabilities from those calculated using $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values alone, due to linkage between groups of titratable residues. Because the protein becomes more highly charged as $p$H is decreased (see Fig. 4), it is natural to expect that the broader distribution of probabilities relative to the Henderson-Hasselbalch approximation may be associated with this increased net charge. Also, there might be more charged residues at the lower $p$H, which might bias the probabilities from Henderson-Hasselbalch values.

However, the Henderson-Hasselbalch probabilities in Fig. 9 already incorporate the influence of the most common charge patterns because they use the top configuration $\alpha_*$ appropriate for each $p$H to construct the needed $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values, via Eq. (10). Thus, the existence of residues that are charged differently at the two $p$H values is not, by itself, sufficient to account for the broader distribution in Fig. 9(d), as compared with that in Fig. 9(b).

Also, for the very top configurations at each $p$H, fewer residues, 43, are modeled as charged at $p$H 4.5 (28 positive, 15 negative, 11 neutral, net charge +13) than at $p$H 7.1, where 48 are charged (25 positive, 23 negative, 6 neutral, net charge +2). Thus, positively and negatively charged residues, taken together, make for a larger total number of charges at $p$H 7.1, despite the fact that the net charge is lower at $p$H 7.1. This situation is physically reasonable because it corresponds mainly to the fact that at $p$H 4.5, eight glutamate and aspartate residues that carried negative charges at $p$H 7.1 are neutral, which can readily occur because the $p$H is much closer to their $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values; H53, H117, G1, and Y174 also change charge. Thus the fact that there are fewer titratable groups that carry charge (positive or negative) at $p$H 4.5 than there are at $p$H 7.1 depends on the set of $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values [see Fig. 15(a) herein

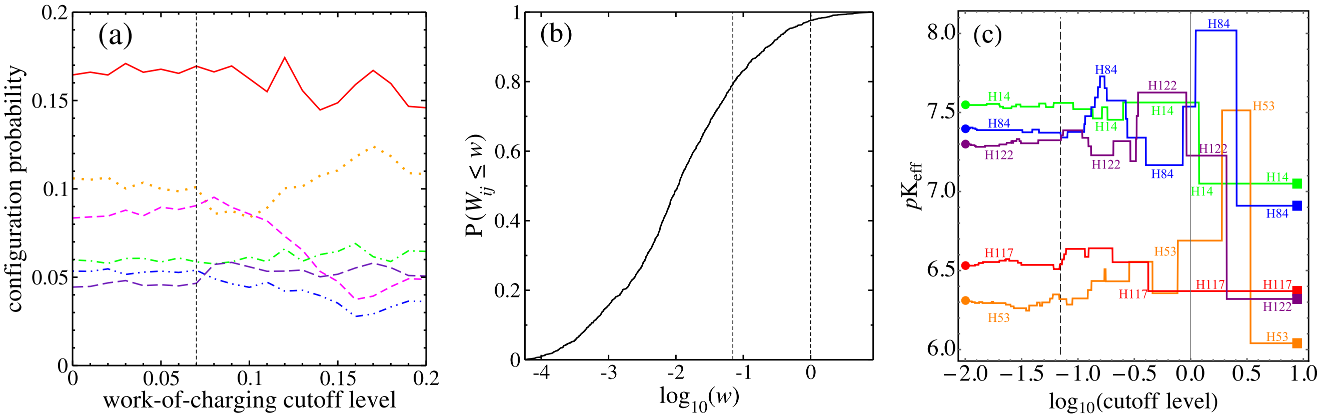

FIG. 10. (a) Dependence on dimensionless work-of-charging cutoff level of the six most common proton configurations at $p$H 7.1 and Debye length 6.0 Å. At each cutoff level on the horizontal axis, all of the entries in $W$ below the given level were set to zero and the probabilities of the configurations were recalculated using Eq. (5). At this $p$H, the order of the configurations is stable up to a cutoff level of 0.07, which is shown by the vertical dashed line. (b) Cumulative distribution function of the work-of-charging matrix entries (see the text). The left vertical line in (b) corresponds to the cutoff level of 0.07 indicated in (a). The right vertical line is at 1. (c) Changes in $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values of the indicated histidine residues, whose protonation switches produce the top-ranked configurations shown in (a). Above a cutoff of 0.07, the H84 $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ value changes are the primary reasons for the configuration probability changes in (a) (see the text).

and Fig. 5 in the Supplemental Material [122]), combined with the bare charge numbers. This is not directly connected with the fact that the deviation from the Henderson-Hasselbalch distribution is greater at $p$H 4.5.

Rather, Eqs. (11) and (12) indicate that the broader width of the distribution of protonation pattern probabilities must arise from the *switches* $\delta\mathbf{q}_\alpha$ of charge patterns from that of $\alpha_*$ that contribute significantly to the factors $e^{-\delta\mathbf{q}_\alpha^\top W \delta\mathbf{q}_\alpha/2}$. More specifically, the broader width of the protonation pattern probabilities relative to the Henderson-Hasselbalch approximation at $p$H 4.5, as compared with that at $p$H 7.1, is due to Glu and/or Asp residue pairs in the same work-of-charging group [see Fig. 3(c)]. Frequent charge switches of these residues at $p$H 4.5 produce the most probable protonation patterns, while at the same time their work-of-charging linkages bias pattern probabilities away from Henderson-Hasselbalch ones. At $p$H 7.1, histidine residue switches produce the most probable patterns, but because each histidine is in a different work-of-charging group, pattern probabilities more closely track the Henderson-Hasselbalch approximation. Figures 9(c) and 9(f) show log-log plots similar to those in Figs. 7(c) and 8(c) for the top-ranked 1000 configurations. At $p$H 7.1, the deviations cluster along lines parallel to the diagonal line of agreement. At $p$H 4.5 this clustering feature is less clear; the scale was expanded to make it apparent.

The polygons linking protonation patterns shown in Figs. 7 and 8 suggest that the use of effective $p$K values holds both value and danger. If the effective $p$K values were to be considered as fixed, they would not account for the lack of independence of the protonation pattern probabilities that is represented graphically by the distortion of the polygons shown. Nevertheless, as suggested by Figs. 9(c) and 9(f), one might accurately model the probability distributions with use of judicious choices of a *changing* set of base configurations $\alpha_*$ for calculating effective $p$K values according to Eq. (10).

How large do off-diagonal parts of the work-of-charging matrix need to be before they significantly affect the probability distribution of protonation configurations, at a given $p$H? Figure 10 examines the sensitivity of the probabilities of the topmost few configurations to the omission of elements of the work-of-charging matrix that are smaller than chosen cutoff levels. Figure 10(a) shows that at $p$H 7.1, the order of the top-ranked six configurations is stable up to a cutoff level of only 0.07, a level that is shown by the vertical dashed line. Such a dimensionless work-of-charging level corresponds to an electrostatic potential $\phi$, produced at one member of a pair of titratable groups by the other, charged member, of $0.07k_BT/e$, or approximately 2 mV. This rather small value to which the ranking of protonation patterns is sensitive occurs for two principal reasons, which are illustrated in Figs. 10(b) and 10(c), respectively.

First, while many of the entries in $W$ are quite small, there are *many* such entries. To quantify this, Fig. 10(b) shows the cumulative distribution function of the work-of-charging matrix entries. While for $0.07 < W_{ij} < 1$, individual titratable site pairs have relatively little effect on one another, a large number of such pairs occurs; 1133 entries are less than 0.07, 262 entries are between 0.07 and 1, and 36 entries are more than 1.

Second, and more specifically, the cutoff values depend on how close the $p$H is to one or more $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values. At the $p$H illustrated, the titration of histidine residues is modeled to account for the relative prominence of the top-ranked configurations, as discussed above and shown by the polygons in Figs. 7 and 8. Further, the agreement between the Henderson-Hasselbalch probabilities and those of the full model for the topmost 32-vertex polygon, shown in Fig. 8(c), suggests that the changes shown in Fig. 10(a) should correspond to changing $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values.

This is borne out by Fig. 10(c), which shows how the $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values of the five histidines change as the work-of-charging cutoff value is increased from 0.01 to 10, all at a $p$H of 7. The 0.07 level is again shown by the vertical dashed line. At cutoffs lower than 0.07, the $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values show small fluctuations much like those of a random walk, a feature that corresponds to the large number of small work-of-charging entries below this level, shown in Fig. 10(b). The ranking of configuration probabilities [shown in Fig. 10(a)] consequently remains stable until the net result of these fluctuations overcomes the difference between two neighboring $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values. This occurs just beyond the 0.07 cutoff level, when the H84 $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ crosses below that of H122. As a result, the H84 $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ is now closer to the ambient $p$H 7.1 and its deprotonation would now be modeled as more probable than that of H122. In terms of the configuration probability polygon in Fig. 7(a), the H84 segments will now be less positively sloped than those of H122. Such a change corresponds precisely to the fact that the configurations initially ranked 2 and 3 switch their order in Fig. 10(a) just above cutoff level 0.07. It is also consistent with the fact that configurations 5 and 6 also switch their rankings at a very similar cutoff level. Further comparison shows that the prominent migration of the H84 $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ value with increasing cutoff level, shown in Fig. 10(c), is largely responsible for the further configuration ranking changes shown in Fig. 10(a). Finally, Fig. 10(c) shows that at higher cutoff levels, many additional switches occur, until the cutoff level is so high that it is larger than any off-diagonal values.

In summary of the implications of Fig. 10, off-diagonal elements of the work-of-charging matrix that are quite small in the dimensionless units $e\phi/k_BT$ can nevertheless change the ranking of protonation configuration probabilities. Also, as larger and larger off-diagonal elements are set to zero, a random-walk-like migration of $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values provides an approximate accounting for the ranking changes of the top configurations, whose probabilities are well represented by the $p\mathrm{K}_{\mathrm{eff},\alpha_*}$ values at this $p$H.

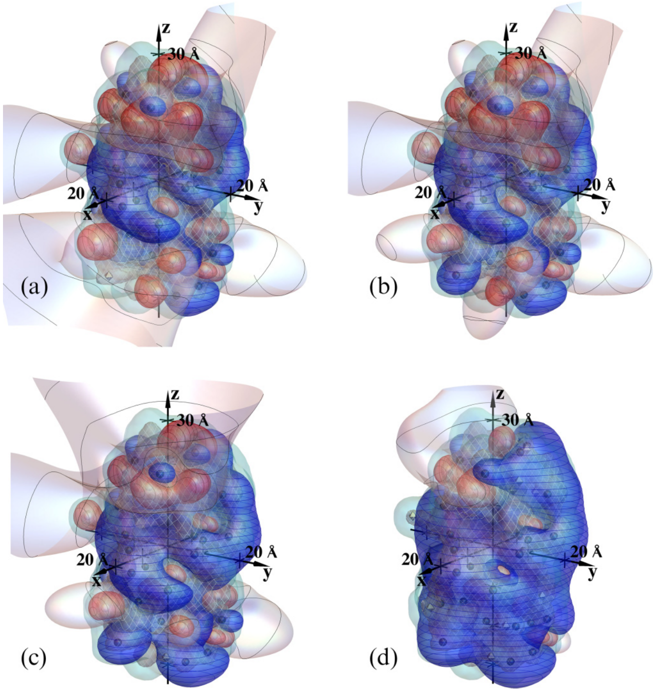

FIG. 11. Screened potential contours of $\gamma$B-crystallin, for the most common protonation patterns at (a) $p\mathrm{H} = 7.1$ (as in Fig. 1), (b) $p\mathrm{H} = 6.5$, (c) $p\mathrm{H} = 5.0$, and (d) $p\mathrm{H} = 4.5$. The Debye length is 6.0 Å; contour values are $+k_BT/e$ V (blue with horizontal curves), 0 V [gray with curves as in Fig. 1(a)], and $-k_BT/e$ V (red with vertical curves); dielectric and electrolyte boundaries are designated as in Fig. 1(a). Note the changing balance between positive and negative voltage regions with $p$H and an accompanying shrinkage of zero-voltage contours, most of which, at $p$H 4.5, extend to less than a Debye length from the protein.

### B. The $p$H dependence of protonation pattern distributions

The modeled probability distributions of protonation configurations show a marked dependence on $p$H, which we now study. By way of introduction, Fig. 11 shows how the screened potential contours change with $p$H for the most common protonation patterns, those occurring at $p\mathrm{H} = 7.1$ [Fig. 11(a), as in Fig. 1(a)], $p\mathrm{H} = 6.5$ [Fig. 11(b)], $p\mathrm{H} = 5.0$ [Fig. 11(c)], and $p\mathrm{H} = 4.5$ [Fig. 11(d)]. The contour values displayed are for $+k_BT/e$ V (blue with horizontal curves), 0 V [gray with curves as in Fig. 1(a)], and $-k_BT/e$ V (red with vertical curves). In each case the Debye length is 6 Å. Prior experimental results, to be analyzed and reported with the help of the model being developed here, led to the choice of $p$H values for Fig. 11. Specifically, at $p$H 7.1, 6.5, and 5.5, at a Debye length of 6 Å, we observe reversible liquid-liquid phase separation in concentrated $\gamma$B-crystallin solutions, strongly suggesting attractive net protein-protein interactions. However, we see no phase separation at $p$H 4.5, and at this $p$H small-angle neutron scattering indicates repulsive interactions.

It is interesting that in this context the balance between the positive and negative voltage regions is fairly even at $p$H 7.1 and $p$H 6.5, while in contrast the positive regions progressively dominate at $p$H 5.0 and $p$H 4.5. Further, the zero potential contours extend far from the protein at the upper three $p$H values shown, but collapse to inside or near the protein at $p$H 4.5. In combination with the findings mentioned above, Fig. 11 suggests that with the more even balance of positive and negative surface regions modeled at the higher $p$H values, neighboring proteins may readily bias their orientations so that oppositely charged surface patches can face one another and interact so as to produce net attractive forces. However, if the balance between positive and negative surface regions becomes skewed beyond that corresponding to Fig. 11(c), net repulsive forces can result. Figure 11(d) illustrates that at $p$H 4.5 the majority of the protein surface is positive. At this $p$H, the angular-averaged interprotein interactions may be relatively insensitive to changes in the particular configuration of protons. It is important to note that a quantitative model will also need to include dispersion forces and hard-core interactions, at least.

Figure 12 shows $\log_{10} P$ vs net protein charge for configurations in the modeled distributions at $p$H 5.5 and $p$H 4.5, together with their single-protonation switch line segments, accompanied by the $p$H 7.1 distribution shown in Fig. 5. As $p$H decreases within this range, there is a substantial spread of net charge and the topmost configuration becomes considerably reduced in probability, reaching below 1 part in a thousand at $p$H 4.5.

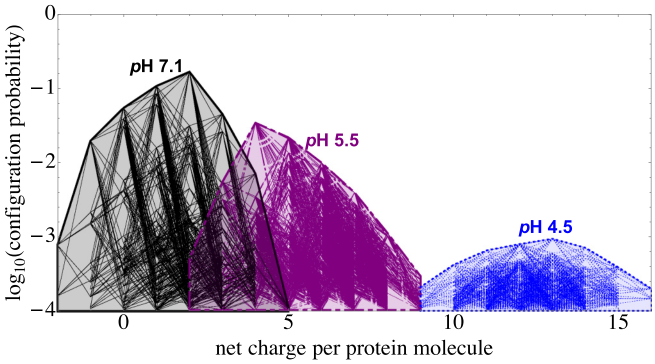

FIG. 12. Plot of $\log_{10} P$ vs net charge, with line segments indicating single-residue protonation switches, for protonation patterns that occur at $p$H 7.1 (black solid line, the same as in Fig. 5), $p$H 5.5 (purple dash-dotted line), and $p$H 4.5 (blue dotted line) (see the text).

Figure 13 illustrates summary statistics of the configuration probability distributions vs $p$H. Figure 13(a) shows that near neutral $p$H the distributions are relatively narrow for this protein; for example, one of the top 100 patterns is expected to occur about 90% of the time. Because $2^7 = 128$, this corresponds to on the order of seven sites switching their protonation status. As discussed in connection with Fig. 5, even though the distribution is relatively narrow near $p$H 7.0, Fig. 13(a) implies that a large number of pairs of patterns may be needed in order to model electrostatically mediated interactions between the proteins. The needed number of pairs can be estimated from the figure. For example, assuming for the purpose of illustration that neighboring patterns do not bias each other's probabilities, it would mean that considering $(100 \times 101)/2$ distinct pairs of patterns would enable one to model a fraction $0.9 \times 0.9$ of the pairs that contribute to the effective interaction strength. The distributions are much broader at lower $p$H values; the $p$H 4.5 curve in Fig. 13(a) shows that at that $p$H, one of the first 1000 patterns will be present only 20% of the time. Figure 13(b) shows the contours of the cumulative probabilities of the top sets of patterns at each $p$H, in the [$p$H, $\log_{10}$(number of configurations)] plane.

Figure 14(a) shows the $p$H dependence of the probabilities of patterns that are each the top pattern within *some* interval of $p$H. To understand these probabilities more thoroughly, consider any pattern $\alpha$ that has a specified number $k_\alpha = n$ of protons bound. Such a pattern has the probability

$$\begin{aligned} P_n &= \frac{\zeta^n e^{-(\boldsymbol{\Delta\mu}^0\cdot\mathbf{O}_\alpha)/k_BT}\; e^{-W_{\mathrm{el},\alpha}/k_BT}}{\mathcal{Q}} \\ &\equiv \frac{\zeta^n B(O_\alpha)}{\mathcal{Q}}, \\ \log_{10} P_n &= \log_{10} B(O_\alpha) - np\mathrm{H} - \log_{10}\mathcal{Q}, \end{aligned} \qquad (14)$$

in which $\zeta = 10^{-p\mathrm{H}}$ and $B(O_\alpha)$ denotes a Boltzmann factor for occupancy vector $\mathbf{O}_\alpha$; $B(O_\alpha)$ includes the intrinsic $p$K values as well as the work-of-charging contribution. Note that all of the $p$H dependence in the last line of Eq. (14) occurs in the last two terms; the partition function $\mathcal{Q}$ in the final term depends on $p$H through $\zeta$. Therefore, for a given value of $n$, all of the curves of $\log_{10} P_n$ vs $p$H are simply vertically displaced with respect to one another, because they differ only due to the quantities $\log_{10} B(O_\alpha)$. This feature is illustrated in Fig. 3(a) in the Supplemental Material [122]. The nearly parabolic shapes in the coordinates ($p$H, $\log_{10} P_n$) correspond to nearly Gaussian shapes when $P_n$ is plotted vs $p$H, as shown in Fig. 14(b) for the top-ranked 12 configurations at $p$H 7.1; these are the configurations illustrated in Fig. 6. In Fig. 14(c) we show the partition function in the form $\log_{10}\mathcal{Q}$ in the range $4 < p\mathrm{H} < 8$. In this $p$H range, $\log_{10}\mathcal{Q}$ can be represented well by cubic or quartic polynomials, specifically $\log_{10}\mathcal{Q} = 474.27 - 86.407 \times p\mathrm{H} + 7.414 \times p\mathrm{H}^2 - 0.30492 \times p\mathrm{H}^3$ (1−adjusted $R^2 = 6 \times 10^{-6}$) and $\log_{10}\mathcal{Q} = 545.64 - 136.56 \times p\mathrm{H} + 20.392 \times p\mathrm{H}^2 - 1.7713 \times p\mathrm{H}^3 + 0.061101 \times p\mathrm{H}^4$ (1−adjusted $R^2 = 5 \times 10^{-7}$), respectively. Such fits can be convenient for estimating $\mathcal{Q}$ in Eq. (5) or (14) for protonation pattern probabilities. The fit residuals in Fig. 14(c) illustrate the degree of error to be expected in using such a fit for $\mathcal{Q}$ and show their polynomial appearance; such correlation of residuals can be detected using, for example, the Durbin-Watson statistic. Physically,

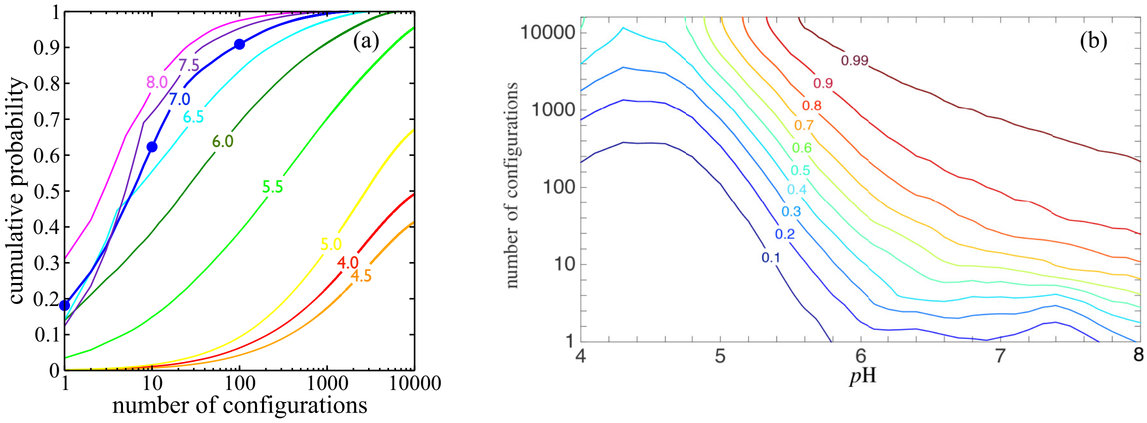

FIG. 13. (a) Cumulative probabilities of the most probable protonation patterns at a Debye length of 6.0 Å. For example, the dots on the $p$H 7.0 curve show that under these conditions, the top $\gamma$B occupancy pattern occurs nearly 20% of the time, one of the first 10 patterns will be present 60% of the time, and one of the top 100 patterns occurs 90% of the time. The $p$H 4.5 curve shows that one of the first 1000 patterns will be present 20% of the time. (b) Contours of the cumulative probabilities displayed in the [pH,$\log_{10}$(number of configurations)] plane. For example, the 0.99 contour indicates that at $p\mathrm{H} \approx 6.8$, 1000 configurations account for 99% of the probability.

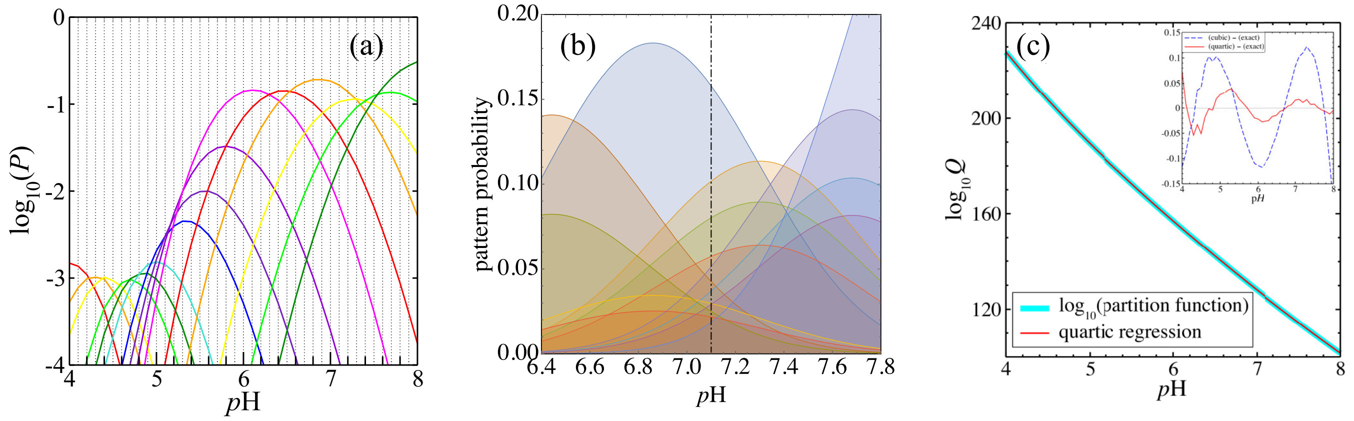

FIG. 14. (a) The $p$H dependence of the probabilities of patterns that are each the top pattern in some interval of $p$H. (b) The $P_\alpha$ vs $p$H for the top 12 configurations at $p$H 7.1 have nearly Gaussian distributions with respect to $p$H. (c) The $\log_{10} \mathcal{Q}$ from the present model, determined using Monte Carlo simulations ($10 \times 10^6$ samples per $p$H, in 0.1 $p$H steps). The inset shows the residuals to two of the fits, which for the quartic fit in this $p$H interval have a range of about 1 part in 4000 of $\log_{10} \mathcal{Q}$.

such an appearance is to be expected given that $\mathcal{Q}$ is a polynomial of essentially higher order than 4, because more than four prominent overall protonation numbers occur in $4 < p\text{H} < 8$ [see Eq. (2) and Fig. 14(a)]. A perspective view of the joint dependence of pattern probabilities on net charge and $p$H is given in Fig. 3(b) in the Supplemental Material [122].

The finding that the $p\text{K}_{\text{eff},\alpha_*}$ values are useful for predicting the ranking of configurations suggests that it is interesting to compare them with the $p\text{K}_{1/2}$ values calculated using the full model. Figure 15 makes such a comparison, with use of $p\text{K}_{\text{eff},\alpha_*}$ values that take $\alpha_*$ to be the top-ranked configuration for $6.6 < p\text{H} < 7.3$ (red closed circles) and to be the top-ranked configuration for $4.4 < p\text{H} < 4.6$ (blue open circles). Figure 15(a) shows that for the histidines that are modeled to titrate near neutral $p$H, the $p\text{K}_{\text{eff},\alpha_*}$ values are indeed almost exactly equal to the corresponding $p\text{K}_{1/2}$ values. This is to be expected from the agreement shown by the black squares in Fig. 8(c). It is instructive to compare the order in which histidine residues first switch to the difference between $p$H 7.1 and their respective $p\text{K}_{\text{eff},\alpha_*}$ values. From Fig. 6 or from Table I, the order of switching is H122, H84, H14, H117, and H53, consistent with the corresponding $|p\text{K}_{\text{eff},\alpha_*} - p\text{H}|$ values of 0.19, 0.29, 0.44, 0.57, and 0.80.

Figure 15(a) also shows that for residues whose $p\text{K}_{\text{eff},\alpha_*}$ values are further from the range for which the chosen $\alpha_*$ configuration is appropriate, the $p\text{K}_{1/2}$ and $p\text{K}_{\text{eff},\alpha_*}$ differ more strongly. Thus, when using $p\text{K}_{\text{eff},\alpha_*}$ as a tool for estimating experimental $p\text{K}_{1/2}$ values, it is important to choose $\alpha_*$ configurations that are prominent, and representative, in a $p$H range that ideally includes the $p\text{K}_{1/2}$ in question. We note that because $1 < p\text{H} < 12$ in the simulations used to create Fig. 15(a), residues with model $p\text{K}_{1/2}$ values outside this range are not shown; in addition to 12 of the arginines and the three cysteines, these included D72 and D107.

Figure 15(b) shows that, except for H53 and H122 below their respective $p\text{K}_{1/2}$ values, the histidine titration curves from the full model agree well with Henderson-Hasselbalch curves, as expected because they are in different work-of-charging groups [Fig. 3(c)]. Thus, although many protonation configuration probabilities are not well predicted by a Henderson-Hasselbalch approach, this may not show up prominently in the titration curves of selected residues.

### C. Dependence of the distributions on ionic strength

The possible effects of ionic strength on solutions of $\gamma$B-crystallin and other proteins are very interesting, in that one expects that they will depend on both the balance and shapes of the negative and positive voltage surface regions. On one hand, if attractive interactions are in part created by protein orientations that put negative and positive surface regions of neighboring proteins face-to-face to some degree, lowering ionic strength would be expected to increase attractions, because then the negative and positive regions would affect one another over a larger range of protein separations. On the other hand, if the net electrostatic portion of the interaction is repulsive, lowering ionic strength would be expected to increase the repulsion. In this context it is interesting to see how large the effects of ionic strength are on the distribution of protonation patterns, which could also play some role in mediating such effects.

Figures 16(a) and 16(b) show that the off-diagonal work-of-charging entries, as expected, can increase substantially as ionic strength is lowered. Figure 16(c) shows the corresponding changes in the configuration distribution at $p$H 7.1. In these coordinates the changes appear modest for the case illustrated. However, the effect is nevertheless evident; it is to make the most prominent configurations slightly more probable, at the expense of some of the less probable configurations. Figure 16(d) shows this in summary fashion. Note the crossover between the changes shown by the top-ranked configurations, whose probabilities generally increase (blue curve above black curve), at the expense of lower-ranked configurations, whose probabilities generally decrease, though not without exception.

### D. Possible implications for protein-protein interactions

As discussed in the Introduction, our primary purpose here is to provide part of a basis for further investigation of the molecular properties that determine the magnitude of

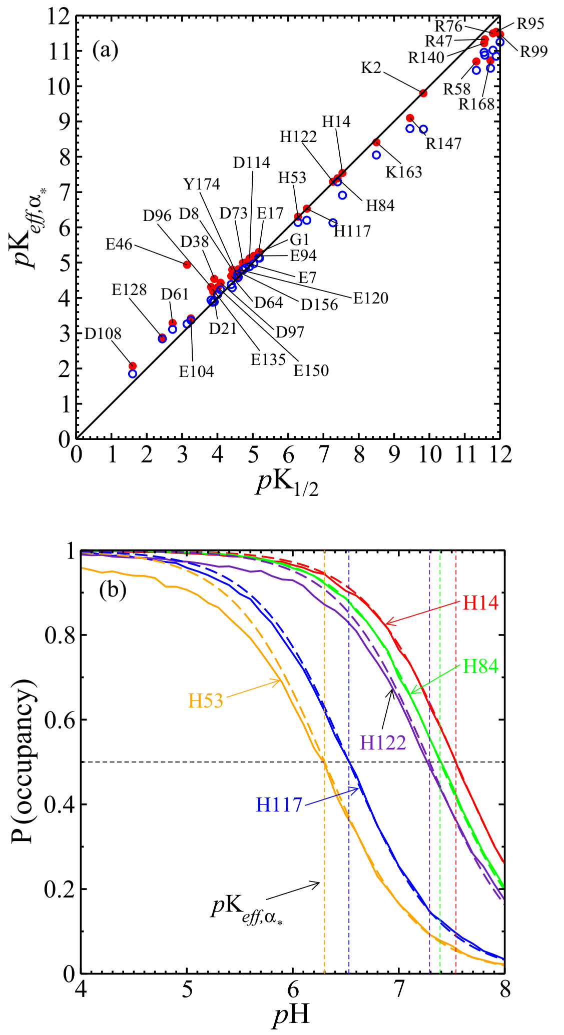

FIG. 15. (a) The $pK_{1/2}$ values from simulations of the full model with interactions, on the horizontal axis, are close to the $pK_{\mathrm{eff},\alpha_*}$ values. Red dots result from taking $\alpha_*$ to be the most prominent protonation configuration within $6.6 < p\mathrm{H} < 7.3$; $pK_{1/2}$ values farther from $6.6 < pK_{1/2} < 7.3$ differ from those $pK_{\mathrm{eff},\alpha_*}$, as expected. Blue open circles result from taking $\alpha_*$ to be the most prominent configuration within $4.4 < p\mathrm{H} < 4.6$; again $pK_{1/2}$ values farther from that of $\alpha_*$ differ more from those of $pK_{\mathrm{eff},\alpha_*}$. (b) Except for H53 and H122 below their respective $pK_{1/2}$, histidine titration curves from the full model (solid line) also agree well with Henderson-Hasselbalch curves (dashed line) as parametrized by the $pK_{\mathrm{eff},\alpha_*}$ values of the $\alpha_*$ used in (a).

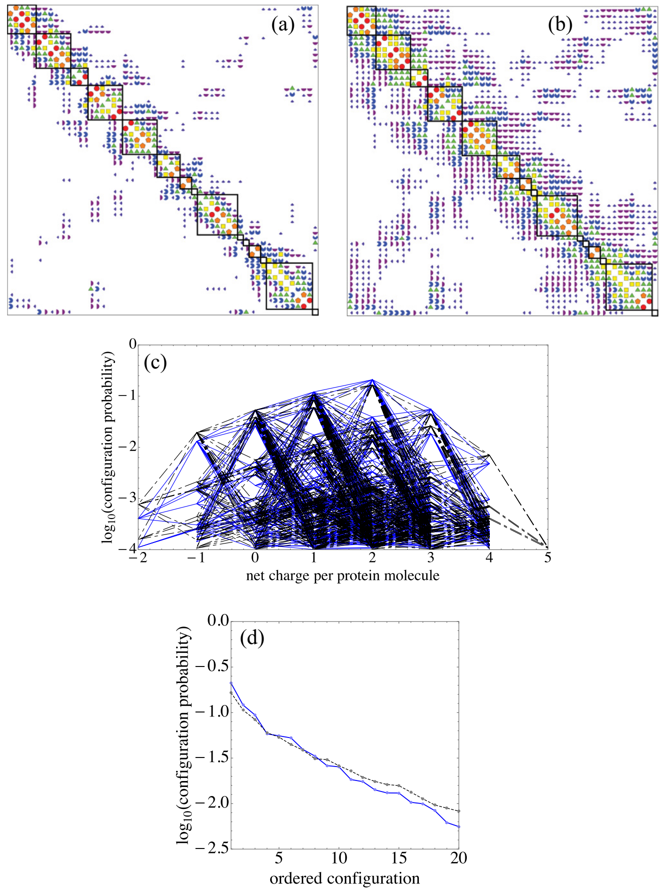

FIG. 16. Lowering ionic strength, corresponding to increasing the Debye length $\lambda_D$ from (a) 6 Å to (b) 20 Å, increases off-diagonal work-of-charging entries and makes the top configurations more prominent while suppressing others, as shown in (c) and (d). In (a) and (b) $W_{ij}$ magnitude codes are as in Fig. 3(a). A 1:1 electrolyte in water at 298 K corresponds to ionic strengths of (a) 257 mM and (b) 23.1 mM. In going from (a) to (b), in all categories but the top (red circles), entries above $0.05k_BT$ increase in number with increased $\lambda_D$. (c) Plot of $\log_{10} P$ vs net charge, as in Fig. 5. Black dash-dotted lines show $\lambda_D = 6$ Å and blue solid lines $\lambda_D = 20$ Å. (d) Plot of $\log_{10} P$ for the top 20 configurations; note the changed vertical scale. The black dashed line shows $\lambda_D = 6$ Å and the blue solid line $\lambda_D = 20$ Å. The contrast between changes in top- and lower-ranked probabilities is shown by the crossing of the dashed and solid curves.

interactions between $\gamma$B- and related $\gamma$-crystallin and other eye lens crystallin proteins, investigation that we hope can eventually achieve sufficient detail to provide for quantitative, predictive modeling of the origin of the cataractogenic effects of single-residue mutations. Because many known cataractogenic mutations of $\gamma$-crystallins involve changes of residue charge, it is natural to study the protonation configuration probability distributions in detail. While many models of orientation-dependent protein-protein interactions have been developed at various levels of coarse graining [3–6,11,132–134], some of which incorporate charge regulation, including models for lysozyme interactions [4–6,132,135,136] and for gamma crystallin interactions [52,88,137], achieving the degree of fine graining for the more predictive modeling needed in many contexts remains an outstanding challenge [19].

Although a quantitative investigation of the consequences of the present model for how site-specific chemical changes influence interactions is not the focus of the present work, we nevertheless comment here on three features that illustrate the scope of the problem. These include (i) the small fraction of the pairs of configurations accounted for by each choice of individual protonation configurations in neighboring proteins, even if those choices are each the top-ranked choice, (ii) the six-dimensional space of the relative positions of two neighboring proteins, and (iii) the biasing of protonation configuration probabilities because of protein proximity [10], a biasing that is itself a function in that same six-dimensional

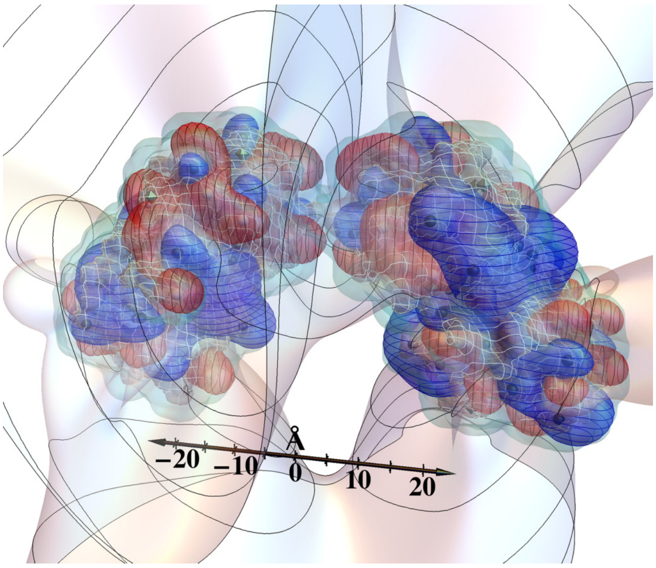

FIG. 17. Screened voltage contours around neighboring $\gamma$B-crystallin molecules, with $p$H 7.1 and Debye length 6 Å. The voltage contour surface designations are as in Fig. 1(a), except that the curves on the 0 V contour surface are spaced by 12 Å from the calculation box center. Each molecule has the top-ranked protonation configuration.

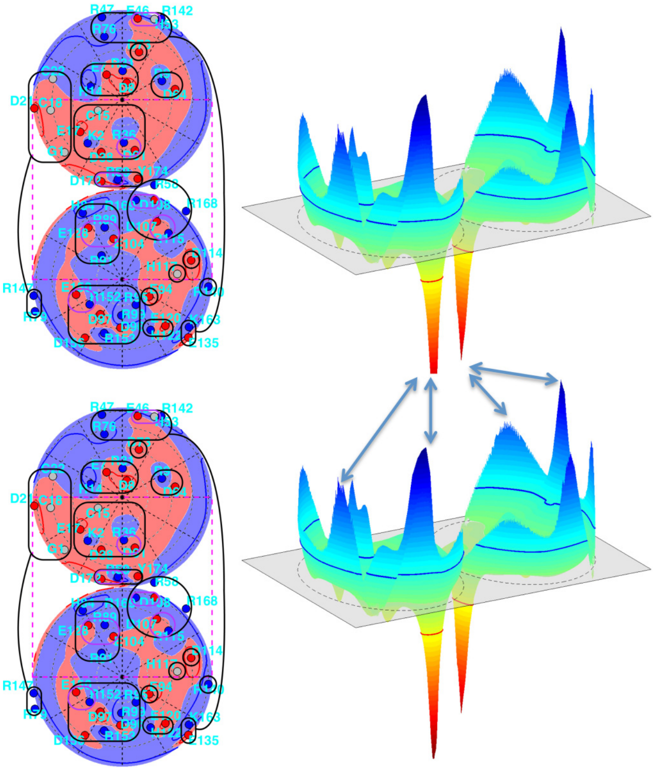

FIG. 18. Visualization of the sets of relative orientations that could give electrostatic attractions between neighboring $\gamma$-crystallins. Lambert projections, as in Fig. 3(c) and described there, of the electrostatic potential at about one-half Debye length from two copies of the most common $p$H 7.1 protonation configuration are shown on the left. Voltages on the same surfaces are plotted vertically on the right, above and below the Lambert projections, with positive up. While the arrows here simply join positive to negative peaks, nonpeak locations can also attract, depending on twist angle.

space. Of course, additional relative position and orientation dimensions are needed if the concentration is high enough so that clusters of more than two neighboring proteins are needed to represent the situation adequately. We now briefly discuss each of these features.

First, Fig. 17 shows the calculated voltage contours around two neighboring $\gamma$B-crystallin molecules, at $p$H 7.1 and Debye length 6 Å. Each of these molecules has been given the most common protonation configuration, that illustrated in Fig. 1. Note that the zero voltage contours appear dramatically altered from those surrounding the isolated protein in Fig. 1(a). Yet the corresponding pair of proton occupancy patterns accounts for only about $0.165 \times 0.165 = 0.0272$ of the contributions to the protein-protein interactions. An illustration of interaction contributions by common pairs of patterns is given in Fig. 1 of the Supplemental Material [122].

Now, for *each* chosen pair of protonation configurations, the space of relative *orientations* of the two proteins has five dimensions, two for each protein to choose the surface points that are in closest proximity and one more for the relative twist about the line joining their centers. Radial separation gives a sixth dimension. Among the many choices of how to visualize the space of possible relative orientations, one shown in Fig. 18 is to make a projection so as to be able to plot the voltages around both protein surfaces above and below two planes and to represent the space of possible proximities by the collection of all line segments or arrows that join pairs of points, one from each plane. Twist can then be added as a position along each line segment, if desired. In Fig. 18, a few such line segments are drawn that indicate connections that *could* correspond to strong electrostatic attractions between neighboring $\gamma$-crystallins, for the most common pair of protonation configurations, shown at left. While the few connections shown in Fig. 18 simply join positive to negative peaks, nonpeak locations can also show electrostatic attractions, depending on the twist angle. With use of calculations that consider the possible relative orientations, one can find prominent basins of attraction and saddle points in the five- or six-dimensional space and illustrate these by points on the appropriate connection lines.

Returning now to Fig. 17, protein proximity dramatically alters the surrounding voltage zero contours, as mentioned above. As a consequence the protonation pattern probability distribution will now reflect between-protein, off-diagonal elements of an enlarged work-of-charging matrix. The existence of these elements means that the joint configuration probabilities can only be approximately represented by the products of probabilities of the individual configurations of hypothetical isolated proteins. As a consequence, the study of the probability distributions of protonation patterns becomes much more intricate for close protein neighbors. Figure 19 gives an example in which changing the relative orientations of two neighboring proteins alters the expanded, two-protein work-of-charging matrix. This example illustrates that the expanded matrices are now functions of the six-dimensional space of the relative positions of the two proteins. Figure 19 also shows that only a small portion of the protein-protein blocks in the expanded matrices differ substantially from zero,

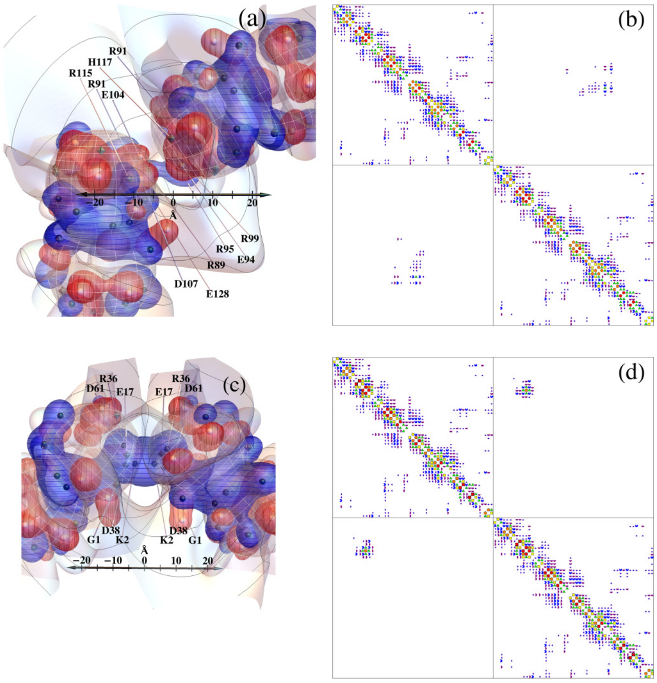

FIG. 19. Changing the relative orientations of two neighboring proteins alters the expanded, two protein work-of-charging matrix. The voltage contour surface designations in (a) and (c) are the same as in Fig. 1(a), except that the curves on the 0 V contour surface are spaced by 12 Å from the calculation box center. The dielectric surfaces are outlined by white longitude and latitude curves. (b) and (d) Same color code as in Fig. 3. (b) and (d) Upper left and lower right squares show the within-protein recalculated work-of-charging matrix entries, while the off-diagonal squares show the between-protein entries. (a) and (c) Different sets of residues are adjacent to one another, as indicated, producing differences in the corresponding work-of-charging matrices shown in (b) and (d), respectively. The order of residues is the same for each protein in (b) and (d), but differs from that in Fig. 3.

which suggests that a perturbation approach might accurately represent the resulting joint probability distributions. To create Fig. 19, we streamlined the needed calculation by using a simpler dielectric boundary than that used above, which consisted of two conjoined, interpenetrating low-dielectric spheres, and by omitting the three cysteine residues that are incorporated in the 54 titratable residues considered above.

Note that close protein proximity can be quite common even at rather low concentrations compared with those that occur in the living eye lens, which can range into the hundreds of milligrams per milliliter. For example, in a square-well model of the phase behavior of $\gamma$B-crystallin [12], Monte Carlo simulations using parameters that gave the closest fit to the observed critical temperature and concentration (a square-well width over diameter of 0.25 and square-well depth of $1.267k_BT$) indicate that even at a concentration of 0.5 mM protein, the mean-field estimate of the average number of contacts per particle [Eq. (30) in [12]] is 0.2; that is, a given protein will already have an essentially close neighbor about 20% of the time. This concentration, which for $\gamma$B-crystallin is close to 10.5 mg/ml, corresponds to a volume fraction of 0.0074, a small fraction of its critical volume fraction of 0.18–0.20, and small compared with estimates of the macromolecular volume fraction in living cells [13], which range from 0.07 to 0.40. Thus one expects altered protonation probability distributions, due to molecular proximity, to contribute substantially to the thermodynamics of protein and other macromolecular solutions within living cells.

## IV. CONCLUSION

We have used the linearized Debye-Hückel approximation to model the probability distributions of protonation patterns on bovine $\gamma$B-crystallin as functions of $p$H. The breadth of the probability distributions indicates that a very large number of pairs of such patterns will be needed in order to account for how the distribution of protonation patterns affects $\gamma$B-$\gamma$B interactions. The key to such an analysis will be to understand not simply the distribution of protonation patterns of a single protein, but rather the probability distribution of protonation patterns, spatial variations of electrostatic potential, and consequent electrostatic interaction free energies present on pairs and larger tuples of neighboring proteins, as functions of their relative positions and orientations.

Accurate, angle-dependent potential of mean force models are needed to provide a sound molecular basis for understanding the statistical thermodynamics and the liquid structure of protein solutions [111–113] and the corresponding dramatic effects of mutations, post-translational modifications, and solution environment on protein phase separation and aggregation in solution. The present model is a step towards building an accurate angle-dependent model of electrostatic contributions to the potential of mean force for $\gamma$-crystallin interactions. Clearly, such a model will also need to encompass other aspects of protein-protein interactions not considered here, including dispersion interactions, the hydrophobic effect, and hydration forces.

## ACKNOWLEDGMENTS

We thank Matthew Lynn for assistance with the use of GAUSSIAN09 and Hassler Thurston for advice about the simulated annealing used to permute the work-of-charging matrix. We thank anonymous referees for their careful readings of the paper, questions, and suggestions that led to clarification. Research reported in this publication was supported by the National Eye Institute of the National Institutes of Health (USA) under Award No. R15EY018249. The content is solely the responsibility of the authors and does not necessarily reflect the official views of the National Institutes of Health (USA).

## APPENDIX: EVALUATION OF ELECTROSTATIC ENERGY INTEGRALS AND CORRESPONDING $p$K SHIFTS

Consider a single charge on the $z$ axis located at $(0,0,z_0)$. We wish to compute $\frac{1}{2}\iiint D \cdot E\, dV$ over the unbounded volume $V$ exterior to a small sphere (neighborhood) of radius $R$, centered at the charge. To do so, we make use of the linearized Poisson-Boltzmann equation

$$\nabla \cdot [\varepsilon(\mathbf{x})\nabla u(\mathbf{x})] = \varepsilon_{\text{out}}\kappa^2(\mathbf{x})u(\mathbf{x}) - \rho(\mathbf{x}).$$

Within $V$, $\rho(\mathbf{x}) = 0$, and in the absence of ionic screening ($\kappa = 0$), the equation reduces to

$$\nabla \cdot [\varepsilon(\mathbf{x})\nabla u(\mathbf{x})] = 0. \tag{A1}$$

To compute the volume integral

$$\frac{1}{2}\iiint_V D \cdot E\, dV = \frac{1}{2}\iiint_V \varepsilon(\mathbf{x})\nabla u(\mathbf{x}) \cdot \nabla u(\mathbf{x}) dV,$$

we use the relation

$$\nabla \cdot [\varepsilon(\mathbf{x})u(\mathbf{x})\nabla u(\mathbf{x})] = \varepsilon\nabla u \cdot \nabla u + u\nabla \cdot [\varepsilon\nabla u].$$

The second term on the right-hand side of the relation is equal to zero by Eq. (A1), in which case we have

$$\varepsilon(\mathbf{x})\nabla u(\mathbf{x}) \cdot \nabla u(\mathbf{x}) = \nabla \cdot [\varepsilon(\mathbf{x})u(\mathbf{x})\nabla u(\mathbf{x})].$$

Then the integral can be written as

$$\frac{1}{2}\iiint_V D \cdot E\, dV = \frac{1}{2}\iiint_V \nabla \cdot [\varepsilon(\mathbf{x})u(\mathbf{x})\nabla u(\mathbf{x})] dV.$$

The integral can be evaluated using Gauss's divergence theorem by closing the volume with a second concentric sphere of radius $R' > R$ and letting $R'$ approach infinity. Let $V'$ be the volume bounded by the two spheres and let $\partial R$ and $\partial R'$ denote the spherical boundary surfaces of radius $R$ and $R'$, respectively. Then

$$\frac{1}{2}\iiint_{V'} \nabla \cdot [\varepsilon u \nabla u] dV = -\frac{1}{2}\iint_{\partial R} \varepsilon(\mathbf{x})u(\mathbf{x})\nabla u(\mathbf{x}) \cdot \mathbf{n}\, dS + \frac{1}{2}\iint_{\partial R'} \varepsilon(\mathbf{x})u(\mathbf{x})\nabla u(\mathbf{x}) \cdot \mathbf{n}' dS, \tag{A2}$$

where the unit vectors $\mathbf{n}$ and $\mathbf{n}'$ are normal to the respective boundary surfaces. We choose the unit normal vectors to be directed outward relative to the spheres, rather than to the volume itself. The second integral on the right-hand side of Eq. (A2) approaches zero as $R'$ approaches infinity. In this limit, we are left with

$$\frac{1}{2}\iiint_V D \cdot E\, dV = -\frac{1}{2}\iint_{\partial R} \varepsilon(\mathbf{x})u(\mathbf{x})\nabla u(\mathbf{x}) \cdot \mathbf{n}\, dS.$$

It is convenient to evaluate the integral in spherical coordinates with the origin translated to the charge location at $(0,0,z_0)$. Since the point charge is located on the $z$ axis and $\varepsilon$ is assumed to be symmetric about the $z$ axis, the integrand is independent of the azimuthal angle $\theta$. Therefore, the surface integral reduces to a single-variable integral given by

$$\frac{1}{2}\iiint_V D \cdot E\, dV = -\pi R^2 \int_0^\pi \varepsilon(\phi)u(\phi)\frac{\partial u}{\partial r}(\phi) \sin\phi\, d\phi.$$

Outside a sphere of radius $r_0$, surrounding an isolated charge of magnitude $q$, placed in a dielectric having relative dielectric coefficient $\varepsilon_r$,

$$\frac{1}{2}\iiint_{r>r_0} D \cdot E\, dV = \frac{q^2}{8\pi\varepsilon_0\varepsilon_r r_0}. \tag{A3}$$

Therefore, if we transfer a charged group surrounded by water, which has dielectric coefficient $\varepsilon_w$, into a medium with coefficient $\varepsilon_r$, the work required is

$$\frac{q^2}{8\pi\varepsilon_0\varepsilon_r r_0} - \frac{q^2}{8\pi\varepsilon_0\varepsilon_w r_0} = \frac{q^2}{8\pi\varepsilon_0 r_0}\left(\frac{1}{\varepsilon_r} - \frac{1}{\varepsilon_w}\right). \tag{A4}$$

The corresponding change in the $p$K of such a group, $\Delta p$K, is therefore given by

$$\begin{aligned}\Delta p\mathrm{K}\ln(10) &= \pm\frac{q^2}{8\pi\varepsilon_0 r_0 k_B T}\left(\frac{1}{\varepsilon_r} - \frac{1}{\varepsilon_w}\right)\\ &= \pm\left[\frac{1}{2k_B T}\iiint_{r>r_0} D \cdot E\, dV - \frac{q^2}{8\pi\varepsilon_0\varepsilon_w r_0 k_B T}\right],\end{aligned} \tag{A5}$$

which corresponds to Eq. (6) in the text. Whereas the last substitution may seem superfluous, in view of Eq. (A3), it is the last equality that enables calculation of how $p$K values can be expected to change in a nonuniform (scalar) dielectric environment, such as the one being used for the present model. This is how we have proceeded (except for the histidines) to model $\gamma$B-crystallin's $p$K values.

To understand which sign is to be used in Eq. (A5), it is valuable to recognize that for a lower dielectric than water, that is, $\varepsilon_r < \varepsilon_w$, the right-hand side of Eq. (A4) is positive, corresponding to the fact that one must do work to bury a charge, of either sign, in a low-dielectric environment. Consider first an acidic residue such as glutamic or aspartic acid. In that case, partially surrounding the charge site with a low-dielectric environment will favor the protonated state, which is uncharged, and therefore a lower concentration of protons (higher $p$H) will suffice for protonation. Thus, the $p$K will shift upward and the + sign should be used in Eq. (A5). The opposite is true for basic residues such as lysine, arginine, and histidine, for which a higher concentration of protons will be needed in order to protonate the site.

[1] J. G. Kirkwood and J. B. Shumaker, Forces between protein molecules in solution arising from fluctuations in proton charge and configuration, Proc. Natl. Acad. Sci. USA **38**, 863 (1952).

[2] C. Tanford and J. G. Kirkwood, Theory of protein titration curves. I. General equations for impenetrable spheres, J. Am. Chem. Soc. **79**, 5333 (1957).

[3] M. Lund and B. Jönsson, Charge regulation in biomolecular solution, Q. Rev. Biophys. **46**, 265 (2013).

[4] B. Kim and X. Song, Calculations of the second virial coefficients of protein solutions with an extended fast multipole method, Phys. Rev. E **83**, 011915 (2011).

[5] H. Y. Chan, V. Lankevich, P. G. Vekilov, and V. Lubchenko, Anisotropy of the Coulomb interaction between folded

proteins: Consequences for mesoscopic aggregation of lysozyme, Biophys. J. **102**, 1934 (2012).

[6] L. J. Quang, S. I. Sandler, and A. M. Lenhoff, Anisotropic contributions to protein-protein interactions, J. Chem. Theory Comput. **10**, 835 (2014).

[7] M. Lund and B. Jönsson, On the charge regulation of proteins, Biochemistry **44**, 5722 (2005).

[8] A. C. Mason and J. H. Jensen, Protein-protein binding is often associated with changes in protonation state, Proteins: Struct. Funct. Bioinf. **71**, 81 (2008).

[9] B. Aguilar, R. Anandakrishnan, J. Z. Ruscio, and A. V. Onufriev, Statistics and physical origins of *p*K and ionization state changes upon protein-ligand binding, Biophys. J. **98**, 872 (2010).

[10] D. Hollenbeck, K. M. Martini, A. Langner, A. Harkin, D. S. Ross, and G. M. Thurston, Model for evaluating patterned charge-regulation contributions to electrostatic interactions between low-dielectric spheres, Phys. Rev. E **82**, 031402 (2010).

[11] M. Lund, Electrostatic chameleons in biological systems, J. Am. Chem. Soc. **132**, 17337 (2010).

[12] A. Lomakin, N. Asherie, and G. B. Benedek, Monte Carlo study of phase separation in aqueous protein solutions, J. Chem. Phys. **104**, 1646 (1996).

[13] D. Hall and A. P. Minton, Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges, Biochim. Biophys. Acta **1649**, 127 (2003).

[14] G. B. Benedek, Cataract as a protein condensation disease: The Proctor Lecture, Invest. Ophthalmol. Visual Sci. **38**, 1911 (1997).

[15] J. I. Clark and J. M. Clark, Lens cytoplasmic phase separation, Int. Rev. Cytol. **192**, 171 (1999).

[16] G. H. Pollack, *Cells, Gels and the Engines of Life* (Ebner, Seattle, 2001).

[17] J. D. Gunton, A. Shiryayev, and D. L. Pagan, *Protein Condensation: Kinetic Pathways to Crystallization and Disease* (Cambridge University Press, Cambridge, 2007).

[18] C. D. Keating, Aqueous phase separation as a possible route to compartmentalization of biological molecules, Acc. Chem. Res. **45**, 2114 (2012).

[19] J. J. McManus, P. Charbonneau, E. Zaccarelli, and N. Asherie, The physics of protein self-assembly, Curr. Opin. Colloid Interface Sci. **22**, 73 (2016).

[20] J. Herzfeld and R. W. Briehl, Phase behavior of reversibly polymerizing systems with narrow length distributions, Macromolecules **14**, 397 (1981).

[21] D. Blankschtein, G. M. Thurston, and G. B. Benedek, Phenomenological theory of equilibrium thermodynamic properties and phase-separation of micellar solutions, J. Chem. Phys. **85**, 7268 (1986).

[22] G. Gompper and M. Schick, Lattice model of microemulsions, Phys. Rev. B **41**, 9148 (1990).

[23] M. Kahlweit, R. Strey, and G. Busse, Microemulsions: A qualitative thermodynamic approach, J. Phys. Chem. **94**, 3881 (1990).

[24] P. van der Schoot and M. E. Cates, Growth, static light scattering, and spontaneous ordering of rodlike micelles, Langmuir **10**, 670 (1994).

[25] J. D. Shore and G. M. Thurston, Charge-regulation phase transition on surface lattices of titratable sites adjacent to electrolyte solutions: An analog of the Ising antiferromagnet in a magnetic field, Phys. Rev. E **92**, 062123 (2015).

[26] K. E. Van Holde, W. C. Johnson, and P. S. Ho, *Principles of Physical Biochemistry* (Pearson/Prentice Hall, Upper Saddle River, 2005).

[27] C. Tanford and R. Roxby, Interpretation of protein titration curves. Application to lysozyme, Biochemistry **11**, 2192 (1972).

[28] S. J. Shire, G. I. H. Hanania, and F. R. N. Gurd, Electrostatic effects in myoglobin. Hydrogen ion equilibria in sperm whale ferrimyoglobin, Biochemistry **13**, 2967 (1974).

[29] D. Bashford and M. Karplus, Multiple-site titration curves of proteins: An analysis of exact and approximate methods for their calculation, J. Phys. Chem. **95**, 9556 (1991).

[30] M. Feig, A. Onufriev, M. S. Lee, W. Im, D. A. Case, and C. L. Brooks, Performance comparison of generalized Born and Poisson methods in the calculation of electrostatic solvation energies for protein structures, J. Comput. Chem. **25**, 265 (2004).

[31] J. Wyman and S. J. Gill, *Binding and Linkage: Functional Chemistry of Biological Macromolecules* (University Science Books, Mill Valley, 1990).

[32] A. Onufriev, D. A. Case, and G. M. Ullmann, A novel view of *p*H titration in biomolecules, Biochemistry **40**, 3413 (2001).

[33] S. Lindman, S. Linse, F. A. A. Mulder, and I. Andre, Electrostatic contributions to residue-specific protonation equilibria and proton binding capacitance for a small protein, Biochemistry **45**, 13993 (2006).

[34] M. A. S. Hass and F. A. A. Mulder, Contemporary NMR studies of protein electrostatics, Annu. Rev. Biophys. **44**, 53 (2015).

[35] U. Sharma, R. S. Negin, and J. D. Carbeck, Effects of cooperativity in proton binding on the net charge of proteins in charge ladders, J. Phys. Chem. B **107**, 4653 (2003).

[36] P. M. Biesheuvel, S. Lindhoud, M. A. Cohen Stuart, and R. de Vries, Phase behavior of mixtures of oppositely charged protein nanoparticles at asymmetric charge ratios, Phys. Rev. E **73**, 041408 (2006).

[37] A. Warshel, S. T. Russell, and A. K. Churg, Macroscopic models for studies of electrostatic interactions in proteins: Limitations and applicability, Proc. Natl. Acad. Sci. USA **81**, 4785 (1984).

[38] T. J. You and D. Bashford, Conformation and hydrogen ion titration of proteins: A continuum electrostatic model with conformational flexibility, Biophys. J. **69**, 1721 (1995).

[39] G. Archontis and T. Simonson, Proton binding to proteins: A free-energy component analysis using a dielectric continuum model, Biophys. J. **88**, 3888 (2005).

[40] A. Warshel, P. K. Sharma, M. Kato, and W. W. Parson, Modeling electrostatic effects in proteins, Biochim. Biophys. Acta **1764**, 1647 (2006).

[41] M. R. Gunner, X. Zhu, and M. C. Klein, MCCE analysis of the *p*Kas of introduced buried acids and bases in staphylococcal nuclease, Proteins: Struct. Funct. Bioinf. **79**, 3306 (2011).

[42] S. Polydorides and T. Simonson, Monte Carlo simulations of proteins at constant *p*H with generalized Born solvent, flexible sidechains, and an effective dielectric boundary, J. Comput. Chem. **34**, 2742 (2013).

[43] E. L. Mehler, M. Fuxreiter, I. Simon, and E. Garcia-Moreno, The role of hydrophobic microenvironments in modulating $p$Ka shifts in proteins, Proteins: Struct. Funct. Bioinf. **48**, 283 (2002).

[44] M. A. Porter, J. R. Hall, J. C. Locke, J. H. Jensen, and P. A. Molina, Hydrogen bonding is the prime determinant of carboxyl $p$Ka values at the N-termini of $\alpha$-helices, Proteins: Struct. Funct. Bioinf. **63**, 621 (2006).

[45] W. R. Forsyth, J. M. Antosiewicz, and A. D. Robertson, Empirical relationships between protein structure and carboxyl $p$Ka values in proteins, Proteins: Struct. Funct. Bioinf. **48**, 388 (2002).

[46] M. R. Gunner, M. A. Saleh, E. Cross, A. ud-Doula, and M. Wise, Backbone dipoles generate positive potentials in all proteins: origins and implications of the effect, Biophys. J. **78**, 1126 (2000).

[47] J. J. Miranda, Position-dependent interactions between cysteine residues and the helix dipole, Protein Sci. **12**, 73 (2003).

[48] M. Lund, R. Vácha, and P. Jungwirth, Specific ion binding to macromolecules: Effects of hydrophobicity and ion pairing, Langmuir **24**, 3387 (2008).

[49] J. Horwitz, I. Kabasawa, and J. H. Kinoshita, Conformation of gamma-crystallins of the calf lens: Effects of temperature and denaturing agents, Exp. Eye Res. **25**, 199 (1977).

[50] J. A. Thomson, P. Schurtenberger, G. M. Thurston, and G. B. Benedek, Binary liquid phase separation and critical phenomena in a protein/water solution. Proc. Natl. Acad. Sci. USA **84**, 7079 (1987).

[51] M. L. Broide, C. R. Berland, J. Pande, O. O. Ogun, and G. B. Benedek, Binary-liquid phase separation of lens protein solutions, Proc. Natl. Acad. Sci. USA **88**, 5660 (1991).

[52] A. Lomakin, N. Asherie, and G. B. Benedek, Aeolotopic interactions of globular proteins, Proc. Natl. Acad. Sci. USA **96**, 9465 (1999).

[53] N. Asherie, Protein crystallization and phase diagrams, Methods **34**, 266 (2004).

[54] G. B. Benedek, J. Pande, G. M. Thurston, and J. I. Clark, Theoretical and experimental basis for the inhibition of cataract, Prog. Retinal Eye Res. **18**, 391 (1999).

[55] A. Pande, J. Pande, N. Asherie, A. Lomakin, O. Ogun, J. A. King, N. H. Lubsen, D. Walton, and G. B. Benedek, Molecular basis of a progressive juvenile-onset hereditary cataract, Proc. Natl. Acad. Sci. USA **97**, 1993 (2000).

[56] N. Asherie, J. Pande, A. Pande, J. A. Zarutskie, J. Lomakin, A. Lomakin, O. Ogun, L. J. Stern, J. King, and G. B. Benedek, Enhanced crystallization of the Cys18 to Ser mutant of bovine gammaB crystallin, J. Mol. Biol. **314**, 663 (2001).

[57] A. Pande, O. Annunziata, N. Asherie, O. Ogun, G. B. Benedek, and J. Pande, Decrease in protein solubility and cataract formation caused by the Pro23 to Thr mutation in human gammaD-crystallin. Biochemistry **44**, 2491 (2005).

[58] J. J. McManus, A. Lomakin, O. Ogun, A. Pande, M. Basan, J. Pande, and G. B. Benedek, Altered phase diagram due to a single point mutation in human gammaD-crystallin. Proc. Natl. Acad. Sci. USA **104**, 16856 (2007).

[59] A. Pande, J. Zhang, P. R. Banerjee, S. S. Puttamadappa, A. Shekhtman, and J. Pande, NMR study of the cataract-linked P23T mutant of human gammaD-crystallin shows minor changes in hydrophobic patches that reflect its retrograde solubility. Biochem. Biophys. Res. Commun. **382**, 196 (2009).

[60] P. R. Banerjee, A. Pande, J. Patrosz, G. M. Thurston, and J. Pande, Cataract-associated mutant E107A of human gammaD-crystallin shows increased attraction to alpha-crystallin and enhanced light scattering, Proc. Natl. Acad. Sci. USA **108**, 574 (2011).

[61] J. G. Kirkwood, Theory of solutions of molecules containing widely separated charges with special application to zwitterions, J. Chem. Phys. **2**, 351 (1934).

[62] S. J. Shire, G. I. H. Hanania, and F. R. N. Gurd, Electrostatic effects in myoglobin. Application of the modified Tanford-Kirkwood theory to myoglobins from horse, California grey whale, harbor seal, and California sea lion, Biochemistry **14**, 1352 (1975).

[63] M. Sundd, N. Iverson, B. Ibarra-Molero, J. M. Sanchez-Ruiz, and A. D. Robertson, Electrostatic interactions in ubiquitin: Stabilization of carboxylates by lysine amino groups, Biochemistry **41**, 7586 (2002).

[64] D. Bashford and K. Gerwert, Electrostatic calculations of the $p\mathrm{K}_a$ values of ionizable groups in bacteriorhodopsin, J. Mol. Biol. **224**, 473 (1992).

[65] D. Bashford, An object-oriented programming suite for electrostatic effects in biological molecules: An experience report on the MEAD project, *International Conference on Computing in Object-Oriented Parallel Environments* (Springer, Berlin, 1997), Vol. 1343, pp. 233–240.

[66] L. Li, C. Li, S. Sarkar, J. Zhang, S. Witham, Z. Zhang, L. Wang, N. Smith, M. Petukh, and E. Alexov, DelPhi: A comprehensive suite for DelPhi software and associated resources, BMC Biophys. **5**, 9 (2012).

[67] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, Electrostatics of nanosystems: Application to microtubules and the ribosome, Proc. Natl. Acad. Sci. USA **98**, 10037 (2001).

[68] J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev, H++: A server for estimating $p\mathrm{K}_a$s and adding missing hydrogens to macromolecules, Nucleic Acids Res. **33**, W368 (2005).

[69] P. Schurtenberger, R. A. Chamberlin, G. M. Thurston, J. A. Thomson, and G. B. Benedek, Observation of Critical Phenomena in a Protein-Water Solution. Phys. Rev. Lett. **63**, 2064 (1989).

[70] P. Schurtenberger, R. A. Chamberlin, G. M. Thurston, and J. A. Thomson, Observation of Critical Phenomena in a Protein-Water Solution, Phys. Rev. Lett. **71**, 3395 (1993).

[71] B. M. Fine, J. Pande, A. Lomakin, O. O. Ogun, and G. B. Benedek, Dynamic Critical Phenomena in Aqueous Protein Solutions, Phys. Rev. Lett. **74**, 198 (1995).

[72] B. M. Fine, A. Lomakin, O. O. Ogun, and G. B. Benedek, Static structure factor and collective diffusion of globular proteins in concentrated aqueous solution, J. Chem. Phys. **104**, 326 (1996).

[73] I. Shand-Kovach, Electrostatic properties of phase-separating bovine lens proteins, Ph.D. thesis, MIT, 1992.

[74] C. Slingsby and L. Miller, The reaction of glutathione with the eye-lens protein $\gamma$-crystallin, Biochem. J. **230**, 143 (1985). Note that bovine $\gamma$B-crystallin was termed $\gamma$II-crystallin at that time.

[75] M. J. McDermott, M. A. Gawinowicz-Kolks, R. Chiesa, and A. Spector, The disulfide content of calf $\gamma$-crystallin, Arch. Biochem. Biophys. **262**, 609 (1988). Note that bovine $\gamma$B-crystallin was termed $\gamma$II-crystallin at that time.

[76] D. A. McQuarrie, *Statistical Mechanics* (University Science Books, Mill Valley, 2000).

[77] R. Kjellander, T. Åkesson, B. Jönsson, and S. Marčelja, Double layer interactions in mono- and divalent electrolytes: A comparison of the anisotropic HNC theory and Monte Carlo simulations, J. Chem. Phys. **97**, 1424 (1992).

[78] R. Kjellander and H. Greberg, Mechanisms behind concentration profiles illustrated by charge and concentration distributions around ions in double layers, J. Electroanal. Chem. **450**, 233 (1998).

[79] Y. Burak and D. Andelman, Hydration interactions: Aqueous solvent effects in electric double layers, Phys. Rev. E **62**, 5296 (2000).

[80] J. M. J. Swanson, J. A. Wagoner, N. A. Baker, and J. A. McCammon, Optimizing the Poisson dielectric boundary with explicit solvent forces and energies: Lessons learned with atom-centered dielectric functions, J. Chem. Theory Comput. **3**, 170 (2007).

[81] I. Borukhov, D. Andelman, and H. Orland, Steric Effects in Electrolytes: A Modified Poisson-Boltzmann Equation, Phys. Rev. Lett. **79**, 435 (1997).

[82] L. Lue, N. Zoeller, and D. Blankschtein, Incorporation of nonelectrostatic interactions in the Poisson-Boltzmann equation, Langmuir **15**, 3726 (1999).

[83] D. Ben-Yaakov, D. Andelman, D. Harries, and R. Podgornik, Beyond standard Poisson-Boltzmann theory: Ion-specific interactions in aqueous solutions, J. Phys.: Condens. Matter **21**, 424106 (2009).

[84] M. Boström, D. R. M. Williams, and B. W. Ninham, The influence of ionic dispersion potentials on counterion condensation on polyelectrolytes, J. Phys. Chem. B **106**, 7908 (2002).

[85] M. Boström, F. W. Tavares, B. W. Ninham, and J. M. Prausnitz, Effect of salt identity on the phase diagram for a globular protein in aqueous electrolyte solution, J. Phys. Chem. B **110**, 24757 (2006).

[86] L. Sandberg and O. Edholm, Nonlinear response effects in continuum models of the hydration of ions, J. Chem. Phys. **116**, 2936 (2002).

[87] H. Gong and K. F. Freed, Langevin-Debye Model for Nonlinear Electrostatic Screening of Solvated Ions, Phys. Rev. Lett. **102**, 057603 (2009).

[88] A. Kurut and M. Lund, Solution electrostatics beyond *p*H: A coarse grained approach to ion specific interactions between macromolecules, Faraday Discuss. **160**, 271 (2013).

[89] Y. Y. Sham, Z. T. Chu, and A. Warshel, Consistent calculations of *p*Ka's of ionizable residues in proteins: Semi-microscopic and microscopic approaches, J. Phys. Chem. B **101**, 4458 (1997).

[90] T. Simonson, Dielectric relaxation in proteins: microscopic and macroscopic models, Int. J. Quantum Chem. **73**, 45 (1999).

[91] C. N. Schutz and A. Warshel, What are the dielectric constants of proteins and how to validate electrostatic models? Proteins: Struct. Funct. Genet. **44**, 400 (2001).

[92] T. Simonson, Dielectric relaxation in proteins: the computational perspective, Photosynth. Res. **97**, 21 (2008).

[93] S. C. L. Kamerlin, M. Haranczyk, and A. Warshel, Progress in *ab initio* QM/MM free-energy simulations of electrostatic energies in proteins: Accelerated QM/MM studies of $p\mathrm{K}_a$, redox reactions and solvation free energies, J. Phys. Chem. B **113**, 1253 (2008).

[94] R. Pericet-Camara, G. Papastavrou, S. H. Behrens, and M. Borkovec, Interaction between charged surfaces on the Poisson-Boltzmann level: The constant regulation approximation, J. Phys. Chem. B **108**, 19467 (2004).

[95] M. Uematsu and E. U. Frank, Static dielectric constant of water and steam, J. Phys. Chem. Ref. Data **9**, 1291 (1980).

[96] V. S. Kumaraswamy, P. F. Lindley, C. Slingsby, and I. D. Glover, An eye lens protein-water structure: 1.2 Angstrom resolution structure of $\gamma$B-crystallin at 150 k, Acta Crystallogr. **52**, 611 (1996).

[97] J. M. Israelachvili, *Intermolecular and Surface Forces*, 3rd ed. (Academic, Waltham, 2011).

[98] R. M. C. Dawson, D. C. Elliott, W. H. Elliott, and K. M. Jones, *Data for Biochemical Research*, 3rd ed. (Oxford University Press, Oxford, 1986).

[99] E. Ellenbogen, Dissociation constants of peptides. I. A survey of the effect of optical configuration, J. Am. Chem. Soc. **74**, 5198 (1952).

[100] E. P. Serjeant and B. Dempsey, *Ionisation Constants of Organic Acids in Aqueous Solution*, IUPAC Chemical Data Series No. 23 (Pergamon, New York, 1979).

[101] W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism*, 2nd ed. (Dover, Mineola, 2005), Chap. 6.

[102] J. D. Jackson, *Classical Electrodynamics*, 3rd ed. (Wiley, Danvers, 1998).

[103] T. Simonson, J. Carlsson, and D. A. Case, Proton binding in proteins: $p\mathrm{K}_a$ calculations with explicit and implicit solvent models, J. Am. Chem. Soc. **126**, 4167 (2004).

[104] J. H. Jensen, H. Li, A. D. Robertson, and P. A. Molina, Prediction and rationalization of protein *p*Ka values using QM and QM/MM methods, J. Phys. Chem. A **109**, 6634 (2005).

[105] E. Alexov, E. L. Mehler, N. Baker, A. M. Baptista, Y. Huang, F. Milletti, J. Erik Nielsen, D. Farrell, T. Carstensen, M. H. M. Olsson, J. K. Shen, J. Warwicker, S. Williams, and J. M. Word, Progress in the prediction of $p\mathrm{K}_a$ values in proteins, Proteins: Struct. Funct. Bioinf. **79**, 3260 (2011).

[106] T. Matsui, T. Baba, K. Kamiya, and Y. Shigeta, An accurate density functional theory based estimation of $p\mathrm{K}_a$ values of polar residues combined with experimental data: From amino acids to minimal proteins, Phys. Chem. Chem. Phys. **14**, 4181 (2012).

[107] S. K. Burger, J. Schofield, and P. W. Ayers, Quantum mechanics/molecular mechanics restrained electrostatic potential fitting, J. Phys. Chem. B **117**, 14960 (2013).

[108] G. R. Grimsley, J. M. Scholtz, and C. N. Pace, A summary of the measured *p*K values of the ionizable groups in folded proteins, Protein Sci. **18**, 247 (2009).

[109] T. E. Creighton, *The Biophysical Chemistry of Nucleic Acids & Proteins* (Helvetian, York, 2010).

[110] T. E. Creighton, *Proteins: Structures and Molecular Properties* (Macmillan, New York, 1993).

[111] J. P. Hansen and I. McDonald, *Theory of Simple Liquids*, 3rd ed. (Academic, New York, 2006).

[112] C. G. Gray and K. E. Gubbins, *Theory of Molecular Fluids, Volume 1: Fundamentals*, International Series of Monographs on Chemistry No. 9 (Oxford University Press, Oxford, 1984).

[113] C. G. Gray, K. E. Gubbins, and C. G. Joslin, *Theory of Molecular Fluids, Volume 2: Applications*, International Series of Monographs on Chemistry No. 10 (Oxford University Press, Oxford, 2011).

[114] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox *et al.*, GAUSSIAN09, revision E.01 (Gaussian Inc., Wallingford, CT, 2009).

[115] H. Li, A. D. Robertson, and J. H. Jensen, Very fast empirical prediction and rationalization of protein $p\mathrm{K}_a$ values, Proteins: Struct. Funct. Bioinf. **61**, 704 (2005).

[116] D. C. Bas, D. M. Rogers, and J. H. Jensen, Very fast prediction and rationalization of $p\mathrm{K}_a$ values for protein-ligand complexes, Proteins: Struct. Funct. Bioinf. **73**, 765 (2008).

[117] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen, PROPKA3: Consistent treatment of internal and surface residues in empirical $p\mathrm{K}_a$ predictions, J. Chem. Theory Comput. **7**, 525 (2011).

[118] C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski, and J. H. Jensen, Improved treatment of ligands and coupling effects in empirical calculation and rationalization of $p\mathrm{K}_a$ values, J. Chem. Theory Comput. **7**, 2284 (2011).

[119] G. Roos, N. Foloppe, and J. Messens, Understanding the $p$Ka of redox cysteines: The key role of hydrogen bonding, Antioxid. Redox Signaling **18**, 94 (2013).

[120] H. Bloemendal, W. de Jong, R. Jaenicke, N. H. Lubsen, C. Slingsby, and A. Tardieu, Ageing and vision: Structure, stability and function of lens crystallins, Prog. Biophys. Mol. Biol. **86**, 407 (2004).

[121] A. Pande, D. Gillot, and J. Pande, The cataract-associated R14C mutant of human $\gamma$D-crystallin shows a variety of intermolecular disulfide cross-links: A Raman spectroscopic study, Biochemistry **48**, 4937 (2009).

[122] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.96.032415 for additional figures and tables.

[123] J. Myers, G. Grothaus, S. Narayanan, and A. Onufriev, A simple clustering algorithm can be accurate enough for use in calculations of $p$Ks in macromolecules, Proteins: Struct. Funct. Bioinf. **63**, 928 (2006).

[124] Y. V. Sergeev, Y. N. Chirgadze, S. E. Mylvaganam, H. Driessen, C. Slingsby, and T. L. Blundell, Surface interactions of $\gamma$-crystallins in the crystal medium in relation to their association in the eye lens, Proteins: Struct. Funct. Bioinf. **4**, 137 (1988).

[125] M. Tanokura, 1 H-NMR study on the tautomerism of the imidazole ring of histidine residues: I. Microscopic $p$K values and molar ratios of tautomers in histidine-containing peptides, Biochim. Biophys. Acta **742**, 576 (1983).

[126] T. Simonson and C. L. Brooks, Charge screening and the dielectric constant of proteins: Insights from molecular dynamics, J. Am. Chem. Soc. **118**, 8452 (1996).

[127] D. Bashford, Macroscopic electrostatic models for protonation states in proteins, Front. Biosci. **9**, 1082 (2004).

[128] I. V. Leontyev and A. A. Stuchebrukhov, Dielectric relaxation of cytochrome c oxidase: Comparison of the microscopic and continuum models, J. Chem. Phys. **130**, 085103 (2009).

[129] G. N. Patargias, S. A. Harris, and J. H. Harding, A demonstration of the inhomogeneity of the local dielectric response of proteins by molecular dynamics simulations, J. Chem. Phys. **132**, 235103 (2010).

[130] M. A. S. Hass, M. Ringkjøbing Jensen, and J. J. Led, Probing electric fields in proteins in solution by NMR spectroscopy, Proteins: Struct. Funct. Bioinf. **72**, 333 (2008).

[131] P. Kukic, D. Farrell, L. P. McIntosh, B. García-Moreno E., K. S. Jensen, Z. Toleikis, K. Teilum, and J. E. Nielsen, Protein dielectric constants determined from NMR chemical shift perturbations, J. Am. Chem. Soc. **135**, 16968 (2013).

[132] M. L. Grant, Nonuniform charge effects in protein-protein interactions, J. Phys. Chem. B **105**, 2858 (2001).

[133] W. Li, B. A. Persson, M. Morin, M. A. Behrens, M. Lund, and M. Zackrisson Oskolkova, Charge-induced patchy attractions between proteins, J. Phys. Chem. B **119**, 503 (2015).

[134] N. Adžić and R. Podgornik, Charge regulation in ionic solutions: Thermal fluctuations and Kirkwood-Schumaker interactions, Phys. Rev. E **91**, 022715 (2015).

[135] C. Gögelein, G. Nägele, R. Tuinier, T. Gibaud, A. Stradner, and P. Schurtenberger, A simple patchy colloid model for the phase behavior of lysozyme dispersions, J. Chem. Phys. **129**, 085102 (2008).

[136] A. Kurut, B. A. Persson, T. Åkesson, J. Forsman, and M. Lund, Anisotropic interactions in protein mixtures: Self assembly and phase behavior in aqueous solution, J. Phys. Chem. Lett. **3**, 731 (2012).

[137] M. K. Quinn, N. Gnan, S. James, A. Ninarello, F. Sciortino, E. Zaccarelli, and J. J. McManus, How fluorescent labeling alters the solution behavior of proteins, Phys. Chem. Chem. Phys. **17**, 31177 (2015).